UNIVERSIDAD CARLOS III DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA TÉCNICA DE INFORMÁTICA DE GESTIÓN

PROYECTO FIN DE CARRERA

Similitud semántica entre conceptos de Wikipedia

Autor: Antonio Mejía Sánchez-Bermejo
Tutora: Damaris Fuentes Lorenzo

Febrero 2013

RESUMEN

La similitud semántica entre palabras ha sido objeto de estudio, durante muchos años, dentro del área de la recuperación de información. El cálculo de la similitud semántica es un procedimiento genérico en una gran variedad de aplicaciones en áreas de Computación Lingüística e Inteligencia Artificial. Ejemplo de ello podría ser su uso en tareas de procesamiento de lenguaje natural, desambiguación de palabras, detección y corrección de errores en la escritura (malapropismo), clasificación de textos, etc.

Podemos encontrar diferentes medidas para el cálculo de la similitud semántica. Sin embargo, a pesar de su uso extendido, la mayoría de medidas cuentan con un problema básico: las fuentes usadas para su cálculo, que pueden dividirse en taxonomías (jerarquías) con contenido limitado o de un determinado dominio, o corpus de gran tamaño. En el caso de las taxonomías, la mayoría de medidas propuestas hasta la fecha suelen usar WordNet, una taxonomía de términos en inglés. Si bien WordNet es útil para aplicaciones de información general, carece de conceptos específicos y nombres propios, no está traducida a diferentes idiomas y sus actualizaciones tardan tiempo en ver la luz. A veces se opta por usar taxonomías específicas para un determinado área, surgiendo así medidas que no son independientes del dominio. En el caso de medidas que usan diccionarios, se requiere una colección de textos que, si no son los adecuados, no proporcionan buenos resultados.

Debido a estos problemas, hoy en día se hace necesaria una fuente de información que contenga la mayor cantidad de entidades del mundo real, que cuente con el consenso de una comunidad amplia y que sus actualizaciones se publiquen de una manera más ágil. Una de las fuentes de conocimiento que cumplen con estos requisitos es la actual Wikipedia. Wikipedia es una enciclopedia online multilingüe escrita colaborativamente por voluntarios. Posee entradas de un vasto número de entidades y conceptos, desde los más generales a los más especializados, conteniendo en diciembre de 2012, en su versión inglesa, hasta 4.146.000 artículos. Estas cualidades la han convertido ya en una fuente muy atractiva para la extracción de información por parte de numerosas aplicaciones.

Este proyecto detalla una serie de medidas tradicionales cuya fuente de información es WordNet, y las adapta a Wikipedia, obteniendo resultados similares e incluso mejorando las basadas en Wikipedia en algunos casos, además de beneficiarse de las ventajas de esta fuente de información.

Palabras clave: similitud semántica, Wikipedia, taxonomía, recuperación de información

ABSTRACT

Semantic similarity between words has been studied for many years within the area of information retrieval. The semantic similarity calculation is a generic procedure in a wide variety of applications in areas of Linguistic Computation and Artificial Intelligence. It is used in tasks of natural language processing, disambiguation of words, detecting and correcting errors in writing, text classification, automatic hypertext links, search engines, etc.

We can find a variety of different measures for calculating the semantic similarity. However, despite their widespread use, most measures have a basic problem: the information sources used for their calculation. These sources can be divided into taxonomies (hierarchies) with limited content or elaborated for a certain domain, or large corpus. In the case of taxonomies, the majority of measures proposed so far tend to use the relationships between pairs of words in WordNet, a taxonomy of terms in English. While WordNet is useful for applications of general information, it lacks of specific concepts and proper nouns, it is not translated into different languages and its updates take time to be published. Some measures use taxonomies specific to a particular area, thus obtaining metrics that are not independent of the domain. Measures that use dictionaries and other corpus require a big collection of texts that, if not adequate, do not provide good results.

Because of these problems, it is necessary a source of information that contains as many real-world entities as possible, which has the consensus of a broad community and whose updates are published more quickly. A knowledge source that meets these requirements is the current Wikipedia. Wikipedia is a multilingual online encyclopedia written collaboratively by volunteers. It provides entries of a vast number of entities and concepts, from general to more specific, containing up to 4,416 million articles in its English version (in December, 2012). These qualities have now become a very attractive source for the extraction of information for many applications.

This project details a set of traditional measures whose source is WordNet, and adapts them to Wikipedia, getting similar results and even better when compared with Wikipedia measures, benefiting from the advantages of Wikipedia as well.

Keywords: semantic similarity, Wikipedia, taxonomy, information retrieval.

INDICE GENERAL

INDICE DE FIGURAS

ÍNDICE DE TABLAS

TABLA DE FÓRMULAS

1. INTRODUCCIÓN

En este capítulo se realizará una introducción a los factores que han propiciado la realización del proyecto, el propósito y entorno de desarrollo del mismo y las características que inicialmente se tienen que tener en cuenta para realizar dicho desarrollo. Ésta será una visión general del trabajo que poco a poco y a lo largo de los siguientes capítulos se irá ampliando, entrando en más detalles para aclarar los aspectos más relevantes. De este modo se pretende dar a conocer y comprender cuáles son los pasos que se han ido realizando y de qué manera.

1.1. Descripción del problema

Hay una gran variedad de aplicaciones, dentro del campo de la recuperación de información, que han suscitado un gran interés en los últimos años. Dichas aplicaciones, pertenecientes en su mayoría al área de Computación lingüística e Inteligencia Artificial, utilizan a veces un procedimiento común para su resolución: el cálculo de la similitud semántica entre dos términos. Esta similitud semántica es utilizada para la resolución de múltiples problemas lingüísticos y semánticos en el ámbito computacional. Podríamos enumerar entre otros, la desambiguación de palabras, detección y corrección de errores de escritura (malapropismos), clasificación de textos, obtención automática de enlaces en hipertextos, buscadores, etc.

El uso extendido de la similitud semántica se demuestra en el gran número de medidas existentes utilizadas para su cálculo. Las medidas más relevantes realizadas hasta ahora pueden dividirse en aquellas que utilizan taxonomías o estructuras jerárquicas, aquellas que usan corpus y aquellas que usan resultados de buscadores. También hay medidas que usan una combinación de ambas.

Una de las taxonomías más usadas como fuente de datos es WordNet. WordNet es una colección de términos en inglés desplegados de manera jerárquica. El uso de WordNet da buenos resultados para el cálculo de similitud entre términos de ámbito general. Sin embargo, tiene una serie de limitaciones importantes:

- Carece de términos específicos y nombres propios, algo que limita bastante su uso.

- Tiene una jerarquía para sustantivos y otra distinta para verbos, haciendo más difícil la comparación entre términos de diferente categoría léxica.
- Pese a que se encuentran diferentes versiones traducidas a otros idiomas, su institución original, Princeton, sólo mantiene la versión oficial en inglés.
- Las actualizaciones tardan mucho tiempo en ver la luz.

Aparte de WordNet, en otros casos son utilizadas taxonomías específicas para un determinado área, no pudiendo ser por tanto independientes del dominio utilizado.

En el caso de utilizar diccionarios y otros corpus, se requiere el procesamiento de una gran colección de textos que además, si no son los adecuados, no producen buenos resultados. También suelen ser dependientes del dominio.

El uso de buscadores web también se ha venido utilizando los últimos años, ya que la web puede cubrir todos los conceptos del mundo real. Sin embargo, sus resultados no superan los de las medidas tradicionales con taxonomías.

Dada la gran importancia del cálculo de la similitud semántica para numerosas aplicaciones, observamos que es necesaria una medida de similitud semántica que:

- Sea computacionalmente escalable: Que no tenga que procesar ingentes cantidades de documentos para obtener buenos resultados.
- Sea fiable: Que su fuente de datos, sea cual fuere, esté consensuada por una comunidad fiable.
- Cubra tanto dominios generales como más específicos; para que puedan ser resueltas el mayor número de peticiones o consultas posibles, independientemente del dominio en el que esté situado el cálculo de la similitud.

Como puede verse tras estas características, el principal escoyo a solventar para resolver nuestro problema lo encontramos en la fuente de datos a utilizar. Necesitamos 1) encontrar y procesar una fuente de conocimiento que cubra esos requisitos para, a partir de ella, 2) elaborar una medida de similitud semántica que ofrezca unos resultados que se acerquen a los obtenidos con medidas tradicionales.

1.2. Propósito del proyecto

El objetivo del proyecto es la obtención de una medida de similitud semántica que mejore los resultados obtenidos con las medidas ya existentes.

Para la obtención de dicha medida es necesaria la utilización de una fuente de datos que cumpla con los requisitos mínimos especificados en el apartado anterior. Una de las fuentes de conocimiento que cumple con ellos y que será la que se utilice para la consecución de este proyecto es Wikipedia.

Wikipedia se ha convertido no sólo en una enciclopedia de consulta online, sino que se ha transformado en una fuente de términos fiable y completa, pudiéndose utilizar para diversas tareas dentro del campo del Procesamiento de Lenguaje Natural (PLN), Recuperación de Información (RI), etc. Wikipedia es una enciclopedia online multilingüe escrita colaborativamente por voluntarios. Provee entradas para un extenso número de conceptos del mundo real que abarcan desde lo más general a lo más específico, cuenta con el consenso de una comunidad amplia y, además, se actualiza de una manera ágil y constante.

Algo considerado virtud, como es la amplitud de conceptos que contiene, es también la mayor dificultad con la que nos encontramos. Además, estos conceptos están clasificados bajo la estructura de las categorías a las que pertenecen, y esta estructura no es una jerarquía bien formada. En otras taxonomías utilizadas para el cálculo de la similitud semántica, como WordNet, se relacionan conceptos en caminos únicos porque su jerarquía se basa en herencia simple, lo que simplifica mucho su estructura. La herencia simple en jerarquías se da cuando un hijo pertenece a un solo padre. La complejidad de la jerarquía contenida en Wikipedia viene dada por los múltiples posibles caminos que hay entre cada uno de sus términos, ya que las categorías a las que pertenecen están clasificadas a través de una jerarquía múltiple (una categoría puede tener múltiples padres), asemejándose más a un sistema de etiquetado (*tagging*) que a una jerarquía propiamente dicha.

Finalmente, cabe mencionar que es importante notar la diferencia entre similitud semántica (*semantic similarity*) y relación semántica (*semantic relatedness*). La segunda engloba a la primera, y tiene en cuenta cualquier tipo de relación. Para el caso que nos ocupa, sólo estamos interesados en calcular una medida de semejanza que represente similitud de significado (*similarity of meaning*), sinónimos, y no de otro tipo de relación.

1.3. Definiciones, acrónimos y abreviaturas

El objetivo de este apartado es la definición de palabras clave, acrónimos y abreviaturas que aparecen a lo largo del documento.

1.3.1. Definiciones

Wikipedia

Enciclopedia libre y políglota de la Fundación Wikimedia (una organización sin ánimo de lucro).

WordNet (Princeton)

Base de datos léxica en el idioma inglés. Agrupa las palabras en conjuntos de sinónimos llamados 'synsets', proporcionando definiciones cortas y generales, y almacenando las relaciones semánticas entre estos conjuntos de sinónimos

Corpus

Conjunto amplio y estructurado de textos (hoy en día en general electrónicamente almacenados y procesados). Se utilizan para hacer el análisis estadístico y pruebas de hipótesis, comprobación de hechos o la validación de reglas lingüísticas en un universo específico.

Diccionario

Obra de consulta de palabras o términos que se encuentran ordenados alfabéticamente. De dichas palabras o términos se proporciona su significado, definición, etimología y ortografía.

Glosario

Catálogo de palabras de una misma disciplina, un mismo campo de estudio, definidas o comentadas.

Ontología

Descripción explícita y formal de conceptos de un dominio de discurso (clases) describiendo características del concepto y restricciones sobre éstas.

Taxonomía

Estructura jerárquica que clasifica conceptos.

Jerarquía

Organización por categorías o grados de importancia entre diversos ítems. Cada uno de estos ítems se representa “encima”, “debajo” o al mismo nivel que otro, en forma de árbol.

Web Crawler

Programa informático que recoge y clasifica la información en Internet.

Índice de correlación de Pearson

Índice que mide la relación lineal entre dos variables aleatorias cuantitativas. Puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas.

Coeficiente de Jaccard

Medida que se utiliza para comparar la similitud entre dos muestras en términos de presencia o ausencia de especies entre los diferentes componentes de la comunidad.

Snippets

Extracto de texto de una página web que aparece por cada enlace obtenido en una página de resultados de un buscador web.

Support Vector Machine (SVM)

Conjunto de algoritmos de aprendizaje supervisado, relacionados con problemas de clasificación y regresión.

tf/idf (term frecuency / inverse document frecuency)

Técnica donde la importancia de una palabra en un artículo es proporcional al número de veces que aparece en un artículo (if), e inversamente proporcional al número de artículos que contiene esa palabra en el corpus (idf).

1.3.2. Acrónimos

WBS

Work Breakdown Structure (estructura de descomposición del trabajo)

RBS

Resource Breakdown Structure (estructura de descomposición de recursos)

PLN

Procesamiento de lenguaje natural

RI

Recuperación de información

UML

Unified Modeling Language

LCS

Least Common Subsumer

LDOCE

Longman Dictionary of Contemporary English.

URL

Uniform Resource Locator

SVM

Support Vector Machine

1.4. Metodología utilizada

Para que el software siga unos pasos preestablecidos y el resultado sea más sólido y fácil de mantener se utilizan metodologías de desarrollo. La elección de la metodología es uno de los aspectos estratégicos determinantes para conseguir éxito en un proyecto.

1.4.1. Modelo de procesos

El estándar IEEE 1074-1991 (IEEE, 1992) especifica los procesos del ciclo de vida a seguir en el desarrollo del software. Determina las actividades esenciales, sin ordenar en el tiempo, que se deben aplicar en el desarrollo del sistema, como puede verse en la Ilustración 1. El orden en el tiempo lo define el ciclo de vida, que se verá a continuación. Este estándar se puede adaptar a las necesidades de cada proyecto.

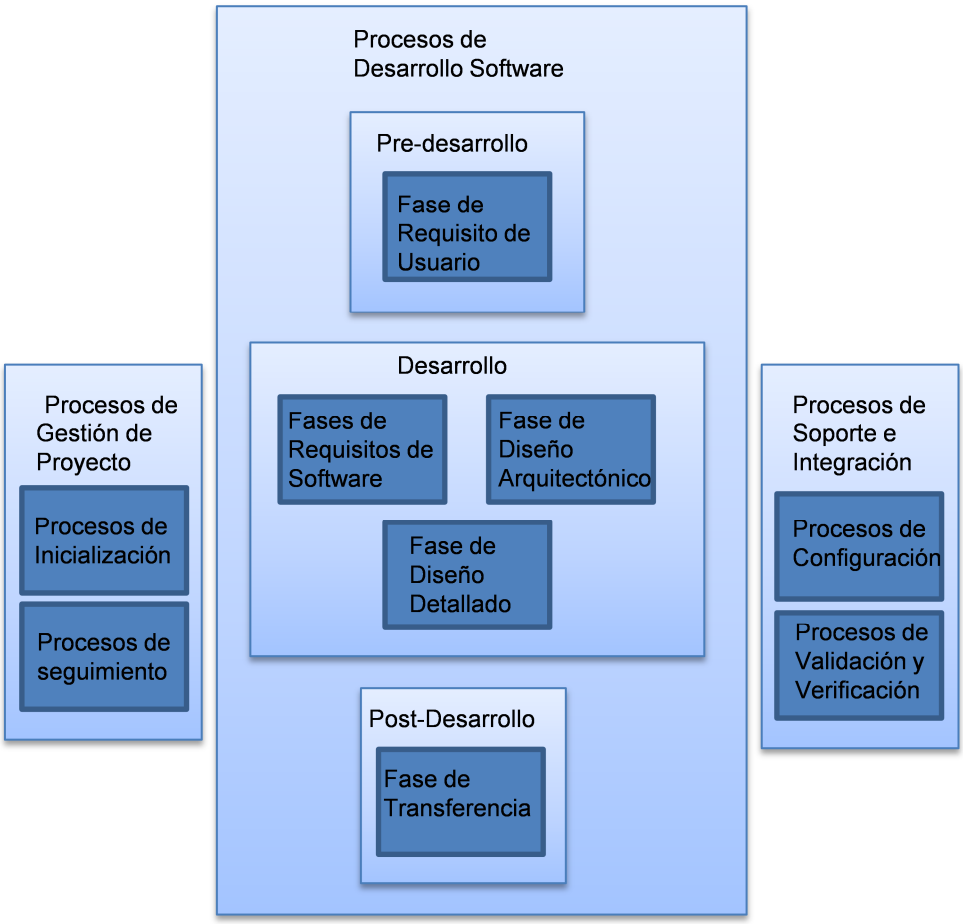

Ilustración 1: Modelo de procesos

En el caso de este proyecto, procesos como el de Configuración o la Fase de Transferencia se omiten directamente.

1.4.2. Ciclo de vida

El desarrollo de este proyecto se guía en base al “ciclo de vida en cascada”, tal y como aparece definido en la ESA (ESA BSSC, 1991). Como puede verse en la Ilustración 2, las fases son ejecutadas de manera secuencial. Cada fase se ejecuta una vez, aunque la iteración de parte de una fase está permitida, siempre que sea para corregir errores o nuevos aspectos. La entrega del sistema ocurre al final, como un solo hito.

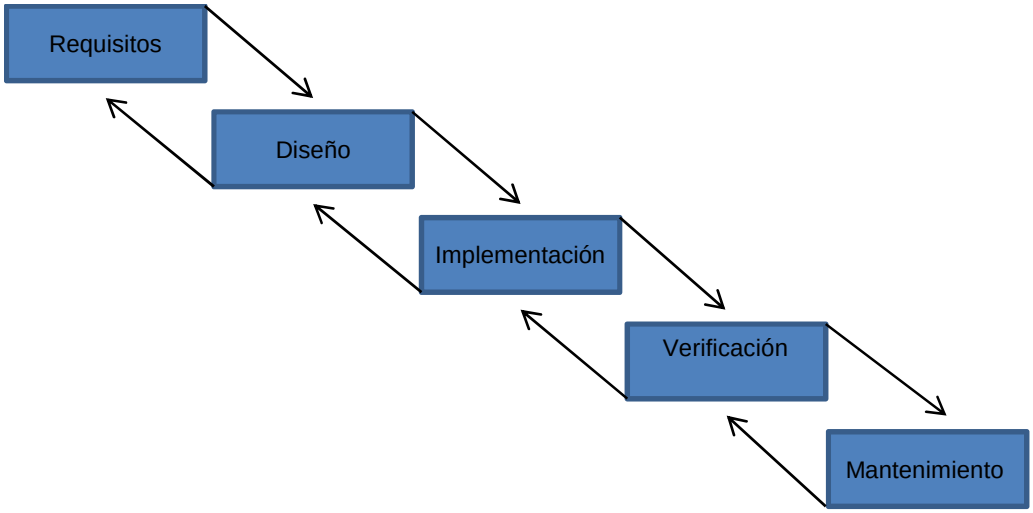

Ilustración 2: Fases del ciclo de vida en cascada

La Ilustración 2 muestra las fases más importantes de este ciclo y las que se han seguido en este proyecto. La fase de “Mantenimiento” no entraría dentro de los objetivos de este proyecto.

1.4.3. Paradigma de modelado

La orientación a objetos será la filosofía a tomar para las siguientes decisiones. En ella se aprecia una proximidad de los conceptos de modelado respecto de las entidades del mundo real, capturando y validando aún mejor los requisitos. Comprende un modelado integrado de propiedades estáticas y dinámicas del ámbito del problema, lo que facilita su construcción, mantenimiento y reutilización. Además, facilita la transición entre distintas fases, favoreciendo el desarrollo iterativo del sistema.

1.5. Notación de modelo utilizada

Para la descripción de los diferentes modelos que aparecerán en los capítulos 4 y posteriores, se utilizará la notación UML (Booch, Rumbaugh, & Jacobson, 2005). UML es un lenguaje estándar de propósito general para el modelado orientado a objetos. Cabe destacar que UML no es una metodología, sino una herramienta gráfica. Los diagramas que en esos capítulos se van a mostrar representan gráficamente partes de un modelo, de una abstracción del proyecto.

1.6. Estructura del documento

El resto de contenidos de este documento está organizado como sigue:

• Capítulo 2: Gestión del Proyecto: Aquí se estudian las actividades en las que se descompondrá el desarrollo del proyecto y su ubicación a lo largo del tiempo, aspectos necesarios para realizar las sucesivas planificaciones. También se analizarán los recursos materiales para su desarrollo. Finalmente, se realizará una comparativa entre el tiempo final planificado y el tiempo real utilizado para la realización del proyecto.

• Capítulo 3: Estado del Arte: Se centra en el estudio del marco sobre el que se encuentra el proyecto. Se hará un repaso de las más importantes medidas de similitud obtenidas hasta la fecha. Estas medidas serán agrupadas por tipo, el cuál vendrá definido por los orígenes de información a partir de los cuales se obtienen las medidas.

• Capítulo 4: Análisis del Problema: Aquí se expone ya de manera clara y directa la problemática del proyecto, explicando de manera breve cuál será nuestra solución a dicho problema.

• Capítulo 5: Diseño general: Partiendo de la problemática expuesta en el capítulo anterior, se pasa a estudiar el diseño de una nueva solución que mejore las ya existentes.

• Capítulo 6: Diseño Detallado: En este capítulo se explica la manera en la que se ha implementado nuestra solución al problema, y qué herramientas y lenguajes se han utilizado.

• Capítulo 7: Adaptación de medidas: Se explica en detalle el proceso llevado a cabo para adaptar las medidas basadas en caminos a Wikipedia.

• Capítulo 8: Experimentos y resultados: Se explica en detalle el procedimiento llevado a cabo para probar las medidas, y se exponen los resultados obtenidos a partir de las mismas.

• Capítulo 9: Conclusiones: Se argumentarán las conclusiones a las que se ha llegado durante la elaboración del proyecto, así como las nuevas ideas a incorporar en un futuro.

• Bibliografía: Aquí se recogen las diferentes referencias a los libros, documentos y direcciones electrónicas de páginas en Internet a las que se ha consultado durante el desarrollo de este proyecto.

• Apéndice A: En este apéndice se recogerán varias tablas con los pares de términos del experimento usados para los experimentos.

2. GESTIÓN DEL PROYECTO

El proyecto que se trata aquí es un producto software y, como tal, se considera como un proceso que consume recursos y está expuesto a condiciones del entorno. Este proceso tiene una finalidad concreta y una serie de objetivos a alcanzar, que deberán ser satisfechos en un plazo determinado, con un coste específico y unos niveles de calidad adecuados. Todos estos aspectos se abarcan en este capítulo.

2.1. Fases

La gestión de un proyecto software comprende las siguientes fases (Amescua, Lopez-Cortijo, & García, 1998):

- Proceso de estimación: Se evaluará el esfuerzo y la duración que harán falta para el desarrollo del proyecto.
- Proceso de organización: Se identifican las tareas a realizar y los recursos necesarios para llevarlas a cabo.
- Proceso de planificación: Se establece el orden en el que se van a realizar las tareas definidas en el proceso anterior.
- Proceso de seguimiento: Permite realizar un seguimiento y controlar el uso de recursos utilizado y el previsto, para ver la evolución real del proyecto.

En la figura siguiente se puede apreciar mejor la relación existente entre los diferentes procesos de gestión.

Ilustración 3: Relación entre los procesos de gestión

El seguimiento puede llevar a una replanificación u obligar a llevar a cabo una nueva estimación. En el desarrollo del proyecto se ha visto la necesidad de una replanificación, como se verá en los siguientes apartados.

2.2. Proceso de estimación

La estimación de proyectos software es una actividad con gran importancia estratégica para la realización de una aplicación. Comprende la evaluación del esfuerzo y duración necesarios para el proyecto, llevándose a cabo en las primeras fases del ciclo de vida. Aun así, una evaluación realizada en momentos tan tempranos del desarrollo implica una gran incertidumbre, por lo que será necesario en este caso ir refinando la estimación inicial a lo largo de todo el proyecto en sucesivos seguimientos.

Con la actividad de estimación se podrá apreciar de forma objetiva y justificable la duración (cronología y tiempo de trabajo) asociada al proyecto, dadas una serie de restricciones.

En este proyecto se seguirá un método de estimación basado en la experiencia. Para ello el tutor, conociendo las estadísticas de proyectos anteriores, prevé una estimación inicial con los siguientes datos:

Tabla 1: Estimación del proyecto

Variable	Valor
Duración	12 meses
Personas	1 persona

Tabla 1 se puede ver que la duración del proyecto estimada inicialmente será de doce meses y sólo será necesaria una persona para realizarlo, que en este caso es el autor del proyecto, que abarcará el desarrollo de las fases de éste. El tutor, aunque es parte fundamental y necesaria para la consecución del proyecto con éxito, no se incluye como persona de trabajo real al limitarse su trabajo a una supervisión, y no a un esfuerzo directo sobre el desarrollo del proyecto.

2.3. Proceso de organización

Este proceso previo a la planificación permite especificar algunos aspectos importantes para la realización de la planificación inicial. Se abarcará la descomposición del proyecto en grupos de actividades elementales y los recursos a utilizar.

2.3.1. Organización de las actividades

La organización de las actividades se recoge en el WBS, que presenta una descomposición de las actividades del proyecto en un diagrama que servirá de soporte para la posterior planificación. Esta lista en forma de árbol es un medio para controlar que no se ha olvidado nada. El diagrama Gantt que se deduzca de este WBS permitirá la planificación del proyecto.

La descomposición inicial de la aplicación queda de la siguiente manera:

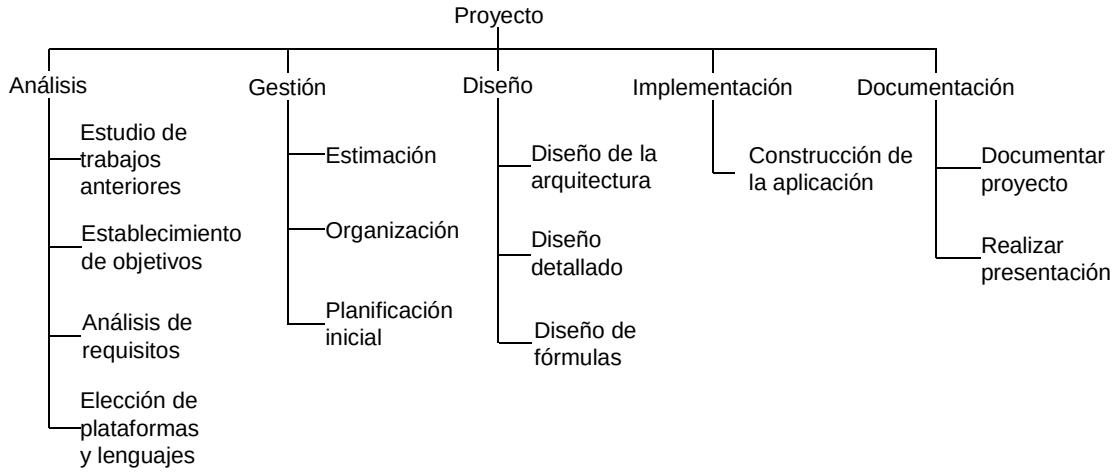

Ilustración 4: WBS inicial

Posteriores avances en el proyecto han hecho que la estructura cambie ligeramente, dando lugar a una descomposición de tareas final que queda como muestra la siguiente ilustración:

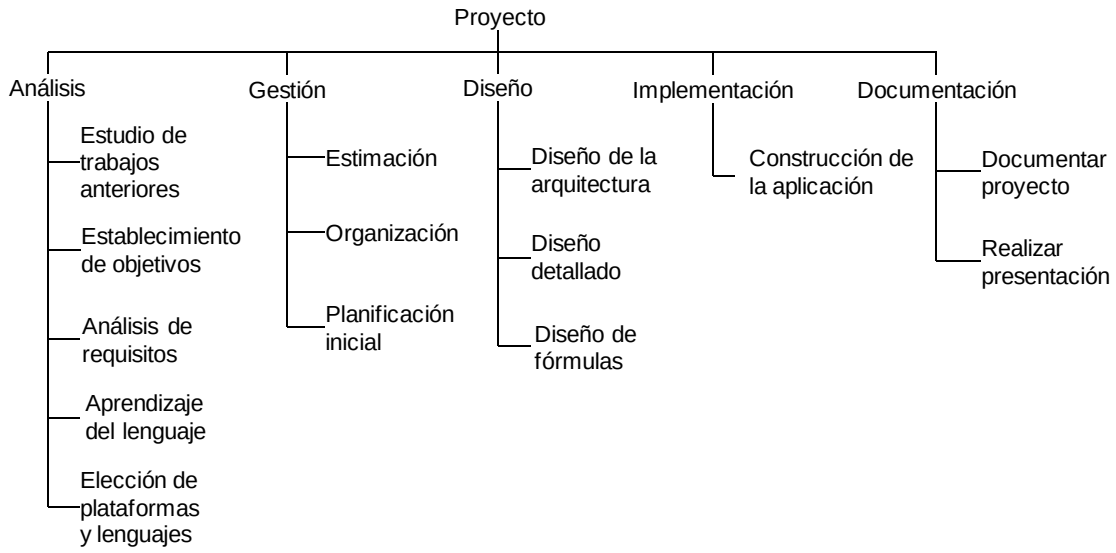

Ilustración 5: WBS final

Como puede verse, tuvo que añadirse una tarea específica para el aprendizaje del lenguaje de implementación, ya que el que se utilizó finalmente, era totalmente desconocido por el autor del proyecto.

2.3.2. Organización de los recursos

Durante el proceso de organización también es necesario representar los recursos materiales y humanos del proyecto. Como los humanos son ya conocidos (1 persona para la realización del proyecto), esta sección se centrará en la representación de los recursos materiales. Para ello se utilizará una forma gráfica denominada RBS, un árbol jerárquico que refleja la estructura de los recursos materiales necesarios para la realización del trabajo (herramientas, plataformas, etc.).

La siguiente figura muestra ese RBS. Se diferencian claramente dos categorías: recursos hardware y recursos software, entre los que destacan las herramientas para la elaboración de la documentación, como Microsoft Office Word 2010 o software de programación para la implementación de este proyecto.

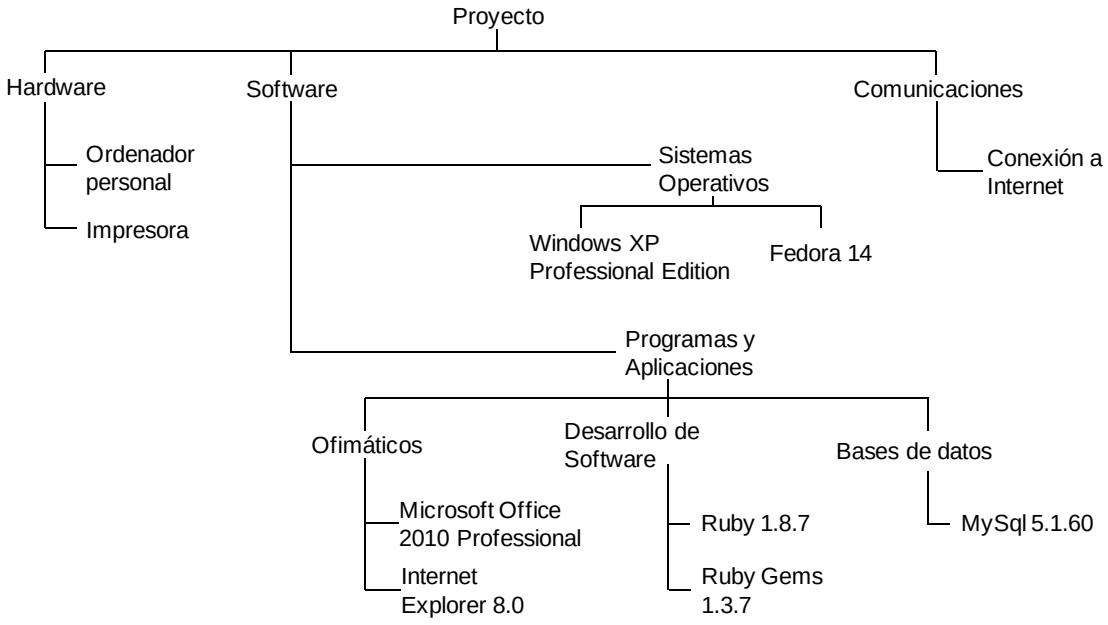

Ilustración 6: RBS final

2.4. Proceso de planificación

Después de haberse especificado cada una de las actividades que deben desarrollarse en la fase de organización, la actividad de planificación establece

en el tiempo dichas actividades, teniendo en cuenta las restricciones existentes, para que así este plan se utilice como referencia a partir de la cual se determine el progreso del proyecto y se corrijan las posibles desviaciones.

Esta actividad, al igual que el resto de actividades de gestión, se realiza en el momento de inicio del proyecto, pero se irá ajustando conforme vaya avanzando su desarrollo. En este apartado se mostrará únicamente la planificación inicial y la que, por diversos motivos, quedó al finalizar el proyecto. Se utilizará para ello las técnicas de los diagramas Gantt, donde se ordenan y programa las tareas teniendo en cuenta también los recursos asignados a cada una de ellas.

Para la realización de la planificación sólo se tendrán en cuenta los datos referidos al tiempo; es decir, no habrá un cálculo de datos de coste. Así pues, la información más relevante es la siguiente:

Tabla 2: Datos relevantes para la planificación

Variable	Valor
Jornada Laboral	2 horas
Semana Laboral	8 horas
Días por mes	16 días

Como puede verse, la jornada laboral de media tiene 2 horas, y durante la semana sólo se trabajan cuatro jornadas (8 horas).

2.4.1. Planificación inicial

Aquí se realiza un primer estudio de las actividades a realizar. La siguiente figura muestra la duración de cada fase y sus fechas estimadas de inicio y final, en función de la fecha prevista de comienzo del proyecto.

	Nombre de la tarea	Fecha de inicio	Fecha de finalización	Duración	Predecesoras
1	Inicio	01/09/11	01/09/11	1	
2	**Análisis**	**02/09/11**	**06/12/11**	**66**	
3	Estudio del contexto	02/09/11	29/11/11	61	1
4	Elección de plataforma y lenguaje	30/11/11	01/12/11	2	3
5	Establecimiento de objetivos	02/12/11	05/12/11	2	4
6	Análisis de requisitos	06/12/11	06/12/11	1	5
7	**Gestión**	**05/09/11**	**07/09/11**	**3**	
8	Estimación	05/09/11	05/09/11	1	
9	Organización	06/09/11	06/09/11	1	8
10	Planificación inicial	07/09/11	07/09/11	1	9
11	**Diseño**	**07/12/11**	**17/02/12**	**50**	
12	Diseño de la arquitectura	07/12/11	14/12/11	6	6
13	Diseño detallado	15/12/11	19/01/12	23	12
14	Diseño de fórmulas	20/01/12	17/02/12	21	13
15	**Implementación**	**20/02/12**	**09/07/12**	**101**	
16	Construcción de la aplicación	20/02/12	09/07/12	101	14
17	**Documentación**	**10/07/12**	**20/09/12**	**53**	
18	Documentar proyecto	10/07/12	10/09/12	45	16
19	Memoria	11/09/12	11/09/12	1	18
20	Realizar presentación	12/09/12	19/09/12	6	19
21	Presentación	20/09/12	20/09/12	1	20
22	Fin	21/09/12	21/09/12	1	21

Ilustración 7: Visión general de la planificación inicial

Para ver mejor la duración de las tareas en el tiempo, se muestra el diagrama Gantt de dicha planificación, en una escala temporal de 1 mes.

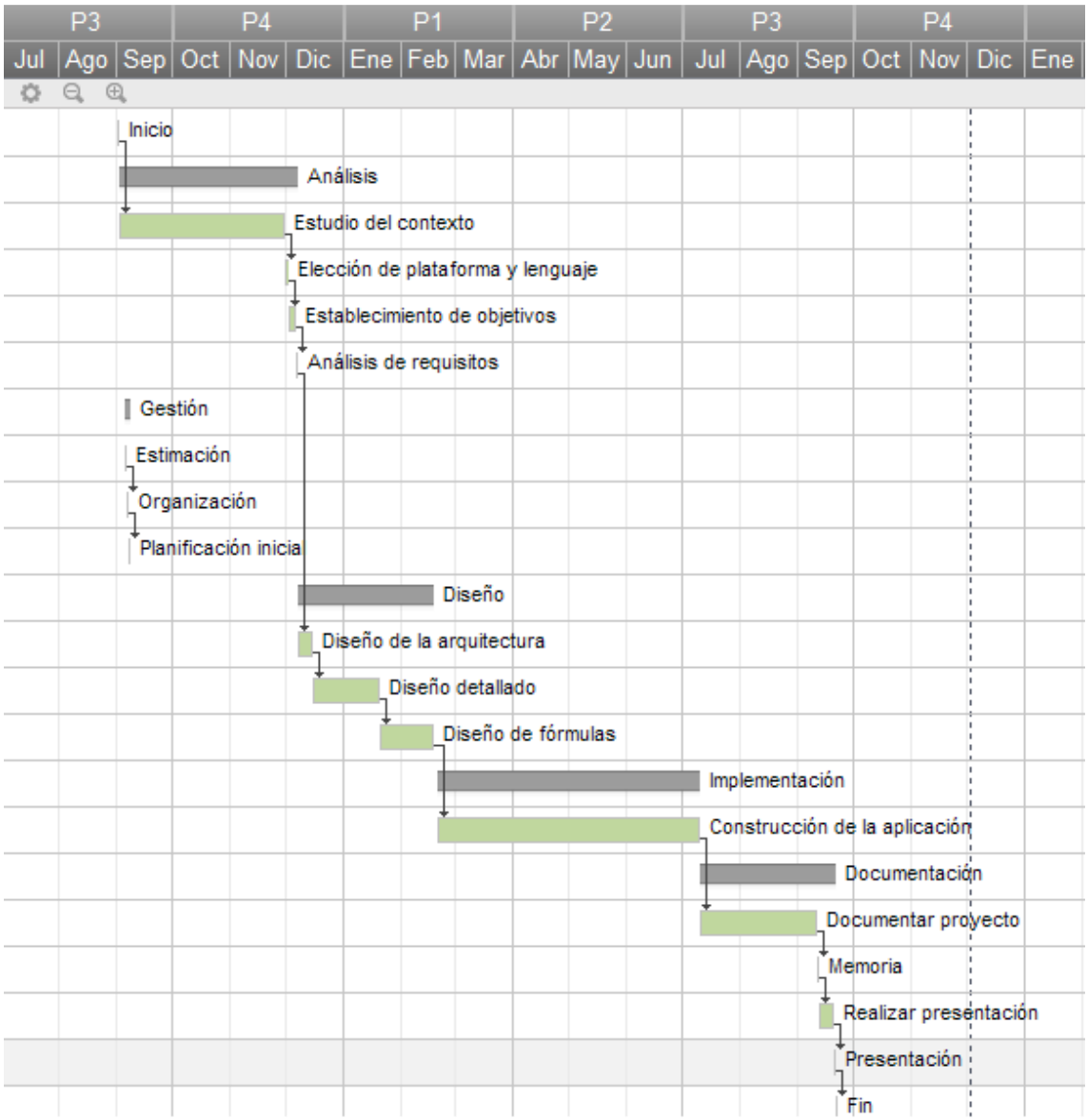

Ilustración 8: Gantt de la planificación inicial

En esta primera planificación el proyecto queda terminado el 21 de septiembre de 2012, habiéndolo iniciado el 1 de septiembre del 2011.

2.4.2. Planificación final

Durante todo el proceso de desarrollo se ha llevado a cabo una planificación constante, con el fin de obtener unos resultados reales del proceso de gestión y, de este modo, poder compararlos con las estimaciones que se habían realizado al comienzo.

Se parte así de las nuevas tareas que se fueron introduciendo o modificando, teniendo los nuevos tiempos que muestra la figura siguiente:

	Nombre de la tarea	Fecha de inicio	Fecha de finalización	Duración	Predecesor
1	Inicio	01/09/11	01/09/11	1	
2	**Análisis**	**02/09/11**	**13/12/11**	**71**	
3	Estudio del contexto	02/09/11	06/12/11	66	1
4	Elección de plataforma y lenguaje	07/12/11	08/12/11	2	3
5	Establecimiento de objetivos	09/12/11	12/12/11	2	4
6	Análisis de requisitos	13/12/11	13/12/11	1	5
7	Aprendizaje de lenguajes	20/09/11	22/11/11	46	
8	**Gestión**	**05/09/11**	**07/09/11**	**3**	
9	Estimación	05/09/11	05/09/11	1	
10	Organización	06/09/11	06/09/11	1	9
11	Planificación inicial	07/09/11	07/09/11	1	10
12	**Diseño**	**14/12/11**	**28/02/12**	**52**	
13	Diseño de la arquitectura	14/12/11	21/12/11	6	6
14	Diseño detallado	22/12/11	26/01/12	23	13
15	Diseño de fórmulas	27/01/12	28/02/12	23	14
16	**Implementación**	**09/11/11**	**12/07/12**	**172**	
17	Construcción de la aplicación	09/11/11	12/07/12	172	
18	**Documentación**	**13/07/12**	**11/10/12**	**65**	
19	Documentar proyecto	13/07/12	01/10/12	57	17
20	Memoria	02/10/12	02/10/12	1	19
21	Realizar presentación	03/10/12	10/10/12	6	20
22	Presentación	11/10/12	11/10/12	1	21
23	Fin	12/10/12	12/10/12	1	22

Ilustración 9: Visión general de la planificación final

Como puede verse, aparece una nueva tarea añadida “Aprendizaje de lenguajes”, quedando las tareas sobre un diagrama Gantt de la siguiente manera:

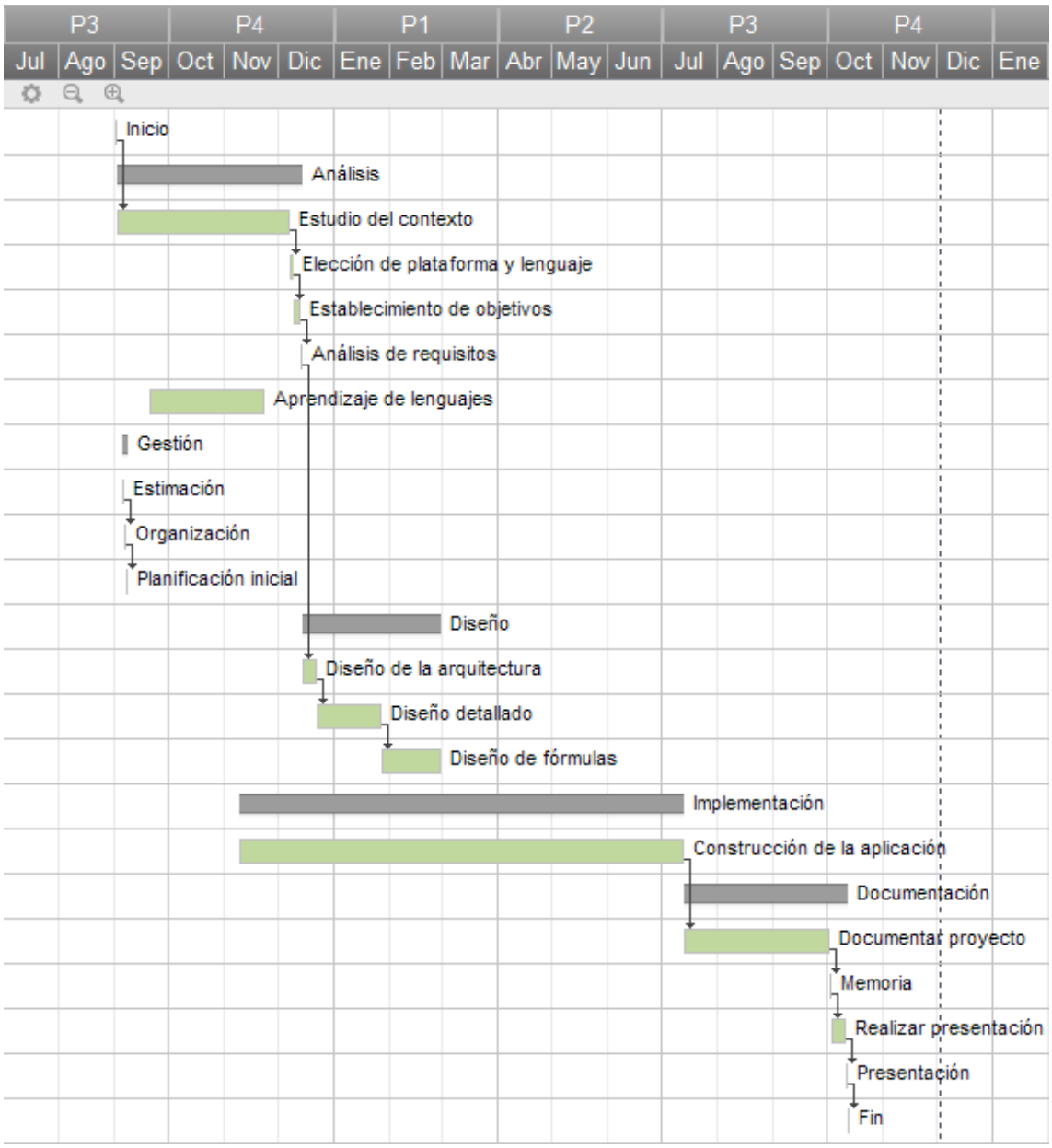

Ilustración 10: Diagrama de Gantt de la planificación final

Puede verse como la tarea “Aprendizaje de lenguajes” se realiza a la vez que el “Estudio del contexto” y con el de “Implementación”, ya que la primera parte de la implementación no necesita de la comprensión del contexto (creación de *crawler* para la descarga y almacenamiento de Wikipedia).

En este caso, la planificación indica que la fecha de terminación final sería el 12 de octubre de 2012.

2.4.3. Comparativa entre el proceso planificado y el real

Una vez terminado el proyecto, se procede aquí a especificar la duración real del mismo, para poder ser luego comparado con la planificación (final) que se había estimado.

En la ilustración se ve que la duración real del proyecto abarca hasta el 24 de diciembre del 2012.

	Nombre de la tarea	Fecha de inicio	Fecha de finalización	Duración	Predecesoras
1	Inicio	01/09/11	01/09/11	1	
2	**Análisis**	**06/08/11**	**13/12/11**	**91**	
3	Estudio del contexto	02/09/11	06/12/11	66	1
4	Elección de plataforma y lenguaje	07/12/11	08/12/11	2	3
5	Establecimiento de objetivos	09/12/11	12/12/11	2	4
6	Análisis de requisitos	13/12/11	13/12/11	1	5
7	Aprendizaje de lenguajes	06/08/11	10/10/11	47	
8	**Gestión**	**05/09/11**	**07/09/11**	**3**	
9	Estimación	05/09/11	05/09/11	1	
10	Organización	06/09/11	06/09/11	1	9
11	Planificación inicial	07/09/11	07/09/11	1	10
12	**Diseño**	14/12/11	29/02/12	53	
13	Diseño de la arquitectura	14/12/11	21/12/11	6	6
14	Diseño detallado	22/12/11	30/01/12	25	13
15	Diseño de fórmulas	31/01/12	29/02/12	22	14
16	**Implementación**	09/11/11	26/07/12	182	
17	Construcción de la aplicación	09/11/11	26/07/12	182	
18	**Documentación**	27/07/12	27/12/12	110	
19	Documentar proyecto	27/07/12	14/12/12	101	17
20	Memoria	17/12/12	17/12/12	1	19
21	Realizar presentación	20/12/12	27/12/12	6	
22	Presentación	21/12/12	21/12/12	1	
23	Fin	24/12/12	24/12/12	1	22

Ilustración 11: Duración real del proyecto

Como puede verse, la duración en el tiempo de algunas tareas ha llegado a ser bastante superior a lo estimado, como en el caso del análisis y de la documentación. El esfuerzo real dedicado al análisis no se ha desviado en gran medida de lo planificado, a excepción de la tarea de documentación, que hace que el esfuerzo total del proyecto se extienda de lo inicialmente planificado. En el siguiente gráfico se puede ver este hecho.

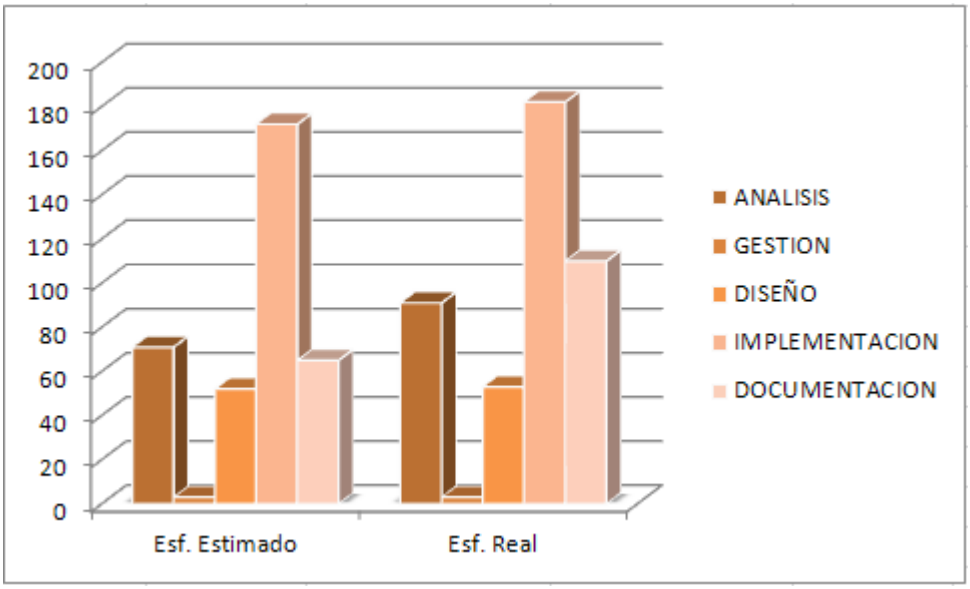

Ilustración 12: Gráfico comparativo entre el esfuerzo estimado y real medido en jornadas

Aquí puede verse que las horas de esfuerzo real para 'Documentación' han sido un poco superiores a las inicialmente estimadas Por otro lado, el resto de fases sí igualan prácticamente las horas estimadas.

2.5. Proceso de seguimiento

Las actividades que se han llevado a cabo en el proceso de seguimiento han consistido en la toma de datos cada semana para el control de cada fase del proceso. De esta manera se recopilaban datos, en concreto de plazos, adaptándose las planificaciones realizadas, analizando desviaciones y realizando acciones correctoras en su caso.

No se ha realizado ningún informe externo de seguimiento. Tras un continuo contacto con la tutora a través del correo, donde se analizaba la evaluación del proyecto y se estudiaban los siguientes pasos a tomar, se realizaba un refinamiento en la planificación del proyecto si se veía necesario, como se ha podido ver en las secciones anteriores.

2.6. Presupuesto estimado

Esta sección ofrece un desglose detallado de los costes y presupuesto estimado del proyecto.

2.6.1. Coste de recursos humanos

Las horas de reuniones se han estimado multiplicando una hora de supervisión por cada semana de duración del proyecto. El resto de horas del autor del proyecto son las obtenidas por la planificación realizada.

Tabla 3: Coste total de los profesionales implicados

Empleado	Horas trabajadas	Coste por hora (€/hora)	Coste total (€)
Contratado1	828(*)	15	12420,00
		Total	12420,00
		Total + 21% IVA	15028,20
(*)387·2+54			

2.6.2. Coste de recursos hardware y software

Los costes de los elementos hardware y software son los siguientes:

Tabla 4: Costes hardware y software

Elemento	Unidades	Tiempo de vida (meses)	Coste unitario (€)	Coste total (€)
Ordenador personal	1	36	400	200,00
Impresora	1	48	150	56,25
Windows XP Professional Edition	1	36	345	172,50
Microsoft Office 2010 Professional	1	24	180	135,00
			Total	563,75
			Total + 21% IVA	682,14

2.6.3. Consumibles

Material de oficina y otros consumibles se muestran a continuación, en la tabla siguiente:

Tabla 5: Costes consumibles

Elemento	Unidades	Coste unitario (€)	Coste total (€)
Material de oficina	-	-	50,00
Toners	1	60	60,00
		Total	110,00
		Total + 21% IVA	133,10

2.6.4. Costes indirectos

Otros costes relacionados con el proyecto son los siguientes:

Tabla 6: Costes indirectos

Elemento	Coste total (€)
Luz	1080,00
Teléfono y conexión	756,00
Total	1836,00
Total + 21% IVA	2221,56

2.6.5. Presupuesto preliminar

El presupuesto preliminar obtenido es el siguiente:

Tabla 7: Presupuesto preliminar

Coste	Coste Total (Con 21% IVA (€))
Recursos Humanos	15028,20
Hardware y software	682,14
Consumibles	133,10
Indirectos	2221,56
Total	18065,00

2.6.6. Posible beneficio

Los beneficios estimados son:

Tabla 8: Beneficio estimado

Elemento	Coste Total + 21%IVA
Beneficio (15% del presupuesto preliminar)	2709,75
Total	2709,75

2.6.7. Presupuesto final estimado

El presupuesto total aplicado al proyecto es pues de veinte mil setecientos setenta y cuatro con setenta y cinco céntimos.

Tabla 9: Presupuesto final

Elemento	Coste Total + 21%IVA
Presupuesto preliminar	18065,00
Beneficio	2709,75
Total	**20774,75**

3. ESTADO DEL ARTE

A partir de la introducción de métodos de computación, se ha intentado obtener durante las últimas décadas un método automatizado, fiable y eficiente para el cálculo de similitud semántica de conceptos. Las medidas resultantes pueden ser utilizadas para una gran variedad de aplicaciones dentro del campo del análisis semántico, que hoy en día están cogiendo gran importancia debido a las inmensas cantidades de información a la que tenemos acceso gracias a Internet.

En este capítulo intentamos presentar el precedente de ideas y medidas previas a nuestro estudio, y que nos servirán para desarrollar nuestro proyecto como una mejora a las posibles carencias o problemas de las medidas ya obtenidas.

3.1. Medidas basadas en corpus

Las medidas basadas en corpus llevan desarrollándose desde la década de los 60, pero los pobres rendimientos computacionales han hecho que hasta la década de los 80 no empezaran a cobrar importancia, debido a que los documentos podían ya ser procesados de manera más eficiente por ordenadores.

Estas medidas se usan sobre todo para medir relaciones semánticas en general, más que para obtener una similitud semántica propiamente dicha. La mayoría de las veces se usan cuando no se cuenta con estructuras jerarquizadas como las que veremos en la sección 3.2.

El primer conjunto de medidas que pueden englobarse en este grupo son las medidas basadas en diccionarios o glosarios *(gloss-based*), que usan las entradas de los términos de un diccionario como fuente de información. El algoritmo de (Lesk, 1986) se puede identificar como el punto de partida para el resurgir de la actividad en este área que continua hasta hoy.

En 1986, Michael Lesk desarrolla un algoritmo para deducir el significado que puede tener una palabra dentro de un contexto. Se basa en la idea de que si dos términos están relacionados, compartirán palabras dentro de sus definiciones. En particular, el algoritmo trata el significado de un término como una bolsa de palabras no ordenadas. De esta manera, para desambiguar un término en una frase, se selecciona el significado de ese término que más palabras comparta con los significados del resto de términos de la frase. Por

ejemplo, queremos desambiguar la palabra *bank* dentro de la frase "*I sat on the bank of the lake"*, y contamos con los siguientes significados:

- *$Bank_1$ = "financial institution that accepts deposits and channels the money into lending activities";*
- *$Bank_2$ = "sloping land especially beside a body of water";*
- *Lake = "a body of water surrounded by land".*

No hay solapamiento entre *Lake* y *$Bank_1$,* pero sí entre *Lake* y *$Bank_2$*, con *body* y *water*, por lo que el algoritmo de (Lesk, 1986) determina que *$Bank_2$* es el significado apropiado para *bank* dentro del contexto dado.

Existe una extensión a este tipo de medidas que se basa en la suposición de que las palabras que se encuentran juntas en la misma frase pueden, de algún modo, estar relacionadas. Teniendo en cuenta este concepto de co-ocurrencia (*co-ocurrence, co-wording*), aparecen medidas basadas en vectores (*vector-based*), las cuales usan la distribución de co-ocurrencias de palabras en diccionarios o grandes corpus (Wilks & et al., 1990); (Church & Hanks, 1990). En estas medidas cada palabra es representada por un vector, donde cada dimensión indica las veces que esta palabra aparece junto con la palabra representada por esa dimensión. La similitud entre conceptos se mide encontrando el coseno (o aplicando otras medidas tradicionales de solapamiento) entre sus respectivos vectores en la matriz de co-ocurrencias. El problema con estos enfoques es que las entradas de los diccionarios suelen ser cortas, y pueden no ofrecer suficiente información sobre la relación entre dos palabras.

Existen otros trabajos, (Kozima & Furigori, 1993), que construyen una *red semántica* de los contenidos de un diccionario. Por cada palabra del diccionario, se crea un nodo, y los enlaces representan la co-ocurrencia de esas palabras en las definiciones. Este enfoque suele ser costoso computacionalmente.

El mayor inconveniente de las medidas basadas en corpus es que se utilizan principalmente para medir relaciones en general, y no medidas de similitud. Además, es básico encontrar el corpus adecuado para obtener buenos resultados, sobre todo cuando nos encontramos con dominios específicos.

3.2. Medidas basadas en un grafo

Para las medidas basadas en grafos, se utilizan taxonomías estructuradas de forma jerárquica. Los conceptos son los nodos dentro de la jerarquía y están unidos por diferentes tipos de relaciones. La más utilizada suele ser la relación de hiperonimia-hiponimia ("es-un"). Por ejemplo, los nodos *manzana* y *fruta* pueden contener una relación de este tipo: *una manzana es una fruta*. Existen otro tipo de relaciones jerárquicas, como la de meronimia-holonimia, que indican que algo es parte o sustancia de otra cosa (*España pertenece a Europa)*; de equivalencia, que se da en términos sinónimos (*respuesta es equivalente a contestación*), etc.

Una de las taxonomías más utilizadas para calcular la similitud o relación semántica entre dos palabras es WordNet. Debido a su alcance cada vez mayor y a la libre disponibilidad, WordNet se ha convertido en un recurso popular para la identificación taxonómica y las relaciones entre conceptos. WordNet contiene información acerca de nombres, verbos, adjetivos y adverbios. Organiza los conceptos relacionados en conjuntos de sinónimos o *synsets*. Cada *synset* representa un concepto o sentido de la palabra. Además de proporcionar estos grupos de sinónimos para representar un concepto, WordNet conecta conceptos a través de relaciones, creándose una red semántica donde los conceptos relacionados pueden ser identificados por su relativa distancia el uno del otro. Las relaciones establecidas son de sinonimia, antonimia, hiperonimia y meronimia. Cada jerarquía, ya sea por nombres o verbos, se puede visualizar como un árbol que tiene conceptos muy generales que se asocian con un nodo raíz y conceptos más específicos.

Una de las primeras medidas basadas en grafos y que utiliza WordNet como taxonomía es la de (Rada et al., 1989). Es una medida simple, ya que evalúa solamente las relaciones de hiperonimia entre los nodos de la taxonomía. De esta manera obtiene la similitud semántica entre dos conceptos calculando la longitud del camino más corto entre los nodos que representan esos conceptos (*shortest path)*. Si una palabra es polisémica puede estar representada por conceptos (nodos) diferentes. En este caso, si no se conoce el sentido de la palabra exacto, tendremos varios caminos; igualmente, se coge el camino más corto.

En la Ilustración 13, podemos ver un extracto de la jerarquía WordNet[1]. Vemos que entre "*dog*" y "*fox*" el camino es "*dog-canine-fox*", es decir, una distancia de 2, y entre "*dog*" y "*cat*" la distancia es de 4, "*dog-canine-carnivore-feline-cat*".

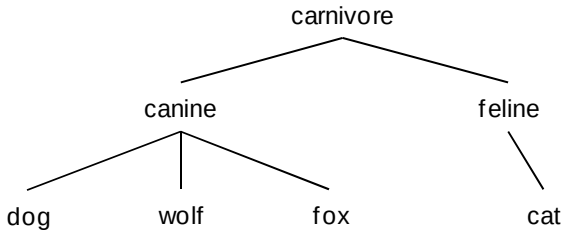

Ilustración 13 : Extracto de la jerarquía WordNet

Los valores de una medida de distancia se encuentran entre 0 (términos idénticos) e infinito. Así es que, para convertirla en una medida de similitud que está entre 0 (ninguna similitud) y 1 (conceptos idénticos) tiene que haber alguna función de transformación. En este caso, la medida de la distancia más corta suele convertirse en medida de similitud restándole a esa distancia el camino más largo posible entre los dos conceptos:

$$sim_{rada}(c_1, c_2) = 2 \cdot d - length(c_1, c_2)$$

Ecuación 1: Medida de similitud de (Rada et al., 1989)

donde *d* es la profundidad máxima de la jerarquía y *length(c_1,c_2)* es el camino más corto entre c_1 y c_2, definido por (Rada et al., 1989).

(Lee & et al., 1993) utiliza esta medida para desarrollar una aplicación que ordena una serie de documentos tras realizar una consulta. Esta ordenación se realiza en función de la similitud de las palabras de la consulta y de las que se encuentran en esos documentos.

El principal problema de este enfoque es que considera que todos los enlaces que relacionan los términos representan distancias uniformes. Es decir, dos términos que se encuentren a la misma distancia en lo alto de la jerarquía tendrán la misma similitud que dos términos que se encuentran en la parte baja de la jerarquía, sin embargo, en la parte alta, los términos son más generales, mientras que en la parte baja son más específicos. Por esto motivo, no podrían considerarse igual de similares dos términos cuyos enlaces se encuentran en distintos niveles de la jerarquía, aunque les separe la misma distancia.

[1] Los extractos citados a lo largo del documento en referencia a WordNet corresponde a la versión 3.1

En el ejemplo de Ilustración 14, podemos comprobar como *person* y *animal* tienen la misma similitud que *boy* y *girl*, al estar a la misma distancia, aun encontrándose en distintos niveles de la jerarquía.

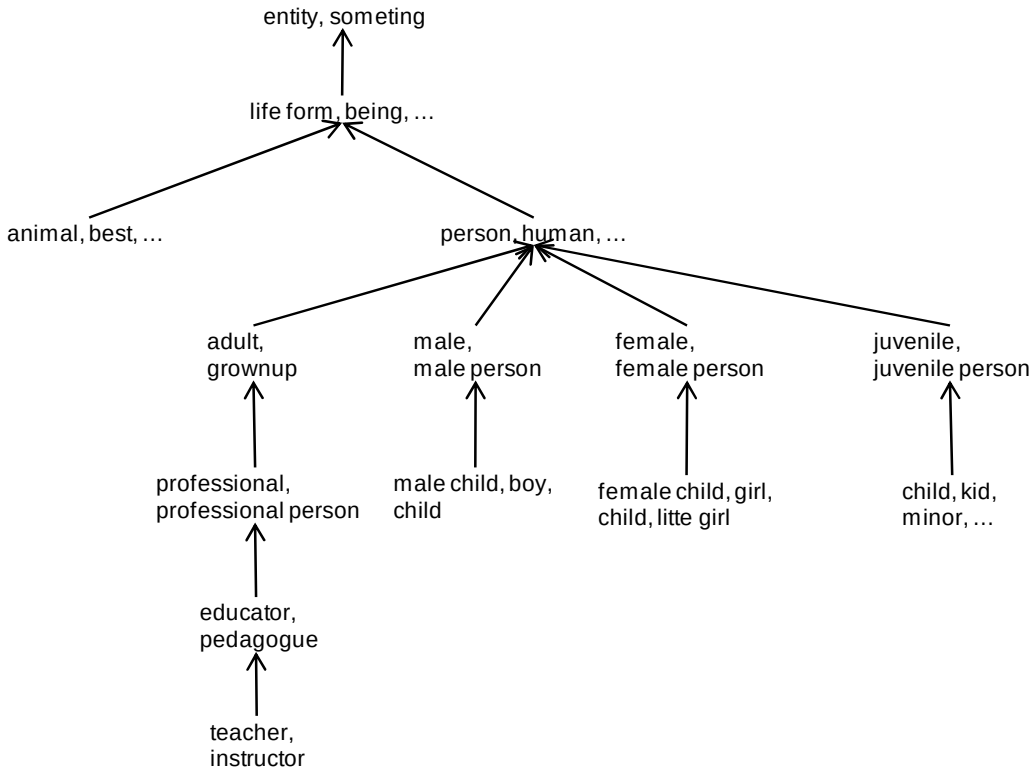

Ilustración 14: Extracto de la jerarquía WordNet

(Sussna, 1993) defiende que los enlaces no son uniformes, por lo que, utilizando WordNet como jerarquía, establece pesos para cada tipo diferente de enlace. Para ello tendrá en cuenta:

- El sitio donde se localiza el enlace. Si está en la parte alta de la jerarquía, puede tener una distancia conceptual mayor que si estuviera en la parte baja, porque la diferenciación se basa en detalles más finos.
- La profundidad de la taxonomía.
- La densidad, que es el número de hijos de un nodo. Cuanto mayor densidad, menor distancia semántica entre los nodos padre e hijo o entre hermanos.
- Tipo de enlace: su medida es de relación en general y no de similitud, pues se vale de todo tipo de relaciones (hiperonimia, meronimia, antonimia, etc).

Otra medida que utiliza la profundidad para el cálculo de la similitud es la de (Leacock & Chodorow, 1994), que usa la longitud mínima del camino entre dos conceptos, pero normalizada a la longitud del camino más largo que puede

haber dentro de la jerarquía. El camino no es exactamente el conteo de arcos en el grafo, sino el de nodos.

$$sim_{lch}(c_1, c_2) = \frac{-\log(length(c_1, c_2))}{2 \cdot d}$$

Ecuación 2: Medida de similitud de (Leacock y Chodorow, 1994)

Otra medida basada también en caminos, pero centrada en la distancia de un nodo a la raíz de la jerarquía es la de (Wu & Palmer, 1994). Calcula la similitud entre dos conceptos c_1 y c_2 teniendo en cuenta la profundidad del concepto c_3 más cercano que incluye a ambos (*least common subsumer)*, al que llamaremos a partir de ahora *lcs*.

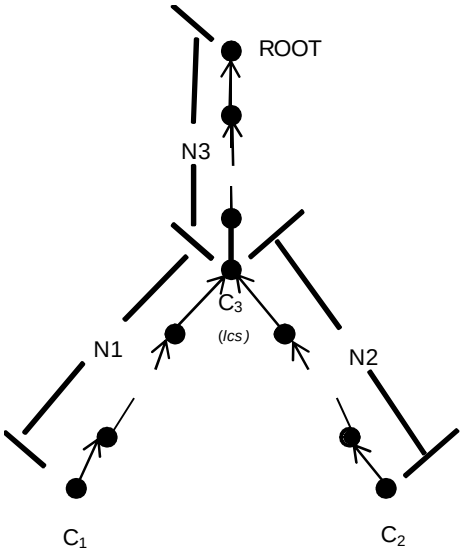

Ilustración 15: Ejemplo ilustrativo de la medida de (Wu y Palmer, 1994)

Para ello, escala la distancia del *lcs* (c_3 en la Ilustración 15) a la raíz, a la suma de las distancias de los conceptos (*N1* y *N2*).

$$Sim_{wp}(c_1, c_2) = \frac{2 * N3}{N1 + N2 + 2 * N3}$$

Ecuación 3: Media de similitud de (Wu y Palmer, 1994)

Se escala para evitar que la medida se base estrictamente en las longitudes de los caminos, que pueden llevar a una medida no exacta, y así evitar el problema de que la similitud entre dos conceptos unidos por un enlace es

diferente según donde nos encontremos en la jerarquía. Como hemos mencionado anteriormente, un enlace entre conceptos muy generales puede implicar una diferencia semántica más grande que dos conceptos más específicos.

Otra propiedad que puede ser utilizada para resolver el problema de uniformidad de enlaces es la densidad conceptual. (Agirre & Rigau, 1996) introducen una medida basada en esta densidad, y la aplican a la desambiguación de nombres. Se centra en las relaciones de hiperonimia en WordNet, y sólo se aplica a la jerarquía de sustantivos.

En la Ilustración 16 se muestra cómo la palabra *W* tiene cuatro significados y varias palabras de su contexto. Cada significado de la palabra W pertenece a una subjerarquía de WordNet. Los puntos representan o el significado de la palabra a desambiguar o las palabras del contexto. Se calcula la densidad conceptual para cada subjerarquía de cada significado. El significado de *W* contenido en la subjerarquía con el valor más alto de densidad conceptual, será el significado no ambiguo de la palabra *W* dentro del contexto dado. En la Ilustración 16 el *Significado 2* será el sentido desambiguado de la palabra *W*.

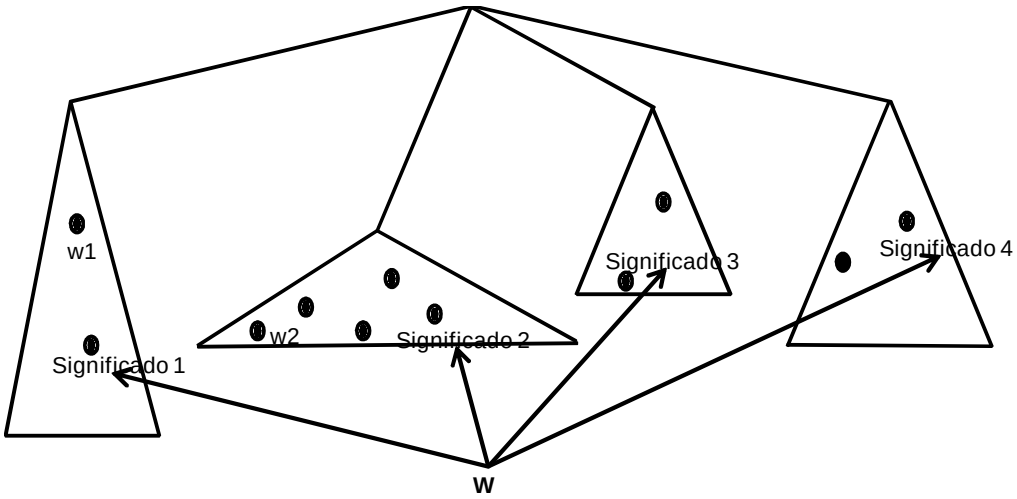

Ilustración 16: Ejemplo ilustrativo de la densidad conceptual de (Agirre y Rigau, 1996)

Otra medida basada en grafos es la de (Blázquez-del-Toro et al., 2008). Se centran en ontologías, aunque se quedan sólo con las relaciones padre e hijo entre clases y, por ello, puede aplicarse a estructuras jerárquicas en general (de hecho, ellos aplican su medida a WordNet). La idea principal es que cuanto más específico es un concepto, menos diferente es de su concepto padre.

Usan la siguiente medida calcular la similitud entre dos conceptos:

$$sim(c_1, c_2) = \max_{lcs \in LCS} \frac{\frac{k\ x\ d_{lcs}}{k\ x\ d_{lcs} + \log\left(\frac{E_{lcs}}{E_{c1}}\right)}\ x\ \frac{k\ x\ d_{lcs}}{k\ x\ d_{lcs} + \log\left(\frac{E_{lcs}}{E_{c2}}\right)}}{\frac{k\ x\ d_{lcs}}{k\ x\ d_{lcs} + \log\left(\frac{E_{lcs}}{E_{c1}}\right)} + \frac{k\ x\ d_{lcs}}{k\ x\ d_{lcs} + \log\left(\frac{E_{lcs}}{E_{c2}}\right)} - \frac{k\ x\ d_{lcs}}{k\ x\ d_{lcs} + \log\left(\frac{E_{lcs}}{E_{c1}}\right)}\ x\ \frac{k\ x\ d_{lcs}}{k\ x\ d_{lcs} + \log\left(\frac{E_{lcs}}{E_{c2}}\right)}}$$

Medida 4: Fórmula de similitud de (Blázquez-del-Toro et al., 2008)

Siendo *k* una constante a calcular, d_{lcs} la profundidad del lcs (que se calcula como la distancia entre el *lcs* y la raíz), y $\left(\frac{E_{lcs}}{E_{c1}}\right)$ o $\left(\frac{E_{lcs}}{E_{c2}}\right)$ el radio de información entre el *lcs* y c_1 o c_2 respectivamente. El cálculo de este cociente se realizará dividiendo la información del *lcs* entre la información del concepto en cuestión.

Para ello, partimos de que un nodo contiene el 100 % de la información de la subjerarquía de la que es raíz, mientras que cada nodo hijo contendrá una fracción equitativa de esta información. En el ejemplo de la Ilustración 17, cada nodo hijo de E_{lcs} (entre ellos, c_{111}), tendrá el 25% de la información total (100/4); a su vez, cada hijo de c_{111} el 5% de ésta (25/5); y, finalmente, c_1 tendrá 1,25 % (5/4).

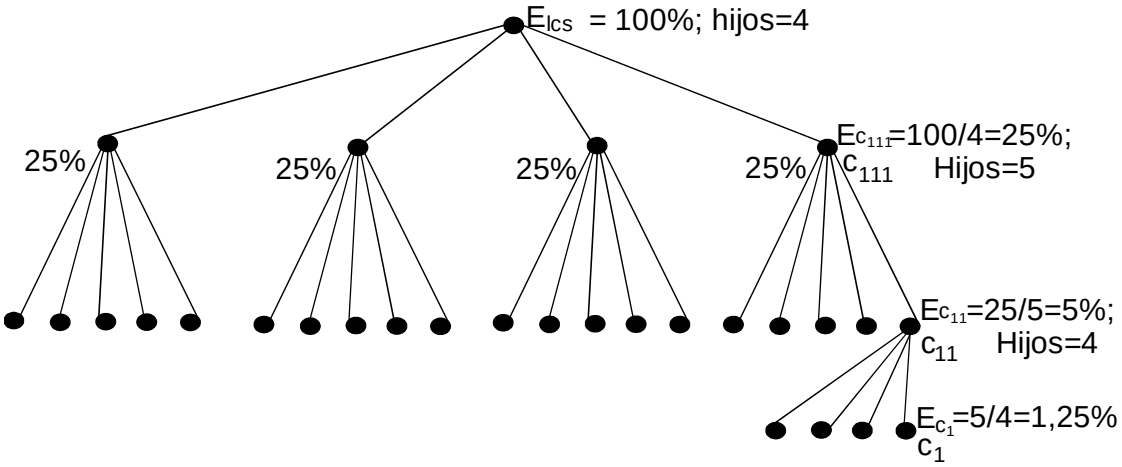

Ilustración 17: Ejemplo ilustrativo para el cálculo del radio de información de (Blázquez-del-Toro et al., 2008)

Tomando de ejemplo de nuevo la Ilustración 17, el cociente $\left(\frac{E_{lcs}}{E_{c1}}\right)$ tendrá como valor

$$\left(\frac{E_{lcs}}{E_{c1}}\right) = \frac{100}{1{,}25} = \frac{100}{100/_{4*5*4}} = 80$$

Ecuación 5: Ejemplo para el cálculo del radio de información de (Blázquez-del-Toro et al., 2008)

Es decir, $\left(\frac{E_{lcs}}{E_{cn}}\right)$ es igual al producto del número de hijos de los padres del concepto c_n hasta llegar al *lcs.*

En general, estos enfoques basados en la distancia entre dos nodos son más intuitivos, sin embargo, siguen dependiendo mucho de la uniformidad entre sus enlaces. Un problema de trabajar con jerarquías como WordNet es que sólo se permite la comparación entre conceptos de la misma categoría léxica (sustantivos con sustantivos, verbos con verbos, etc).

3.3. Medidas basadas en múltiples fuentes de información

Las medidas de este grupo se sirven de un corpus en conjunción con una taxonomía léxica para calcular la similitud semántica entre conceptos. De esta manera, el modelo estadístico de un corpus se ayuda de un espacio conceptual estructurado por esa taxonomía. El factor que introduce la mayoría es *el contenido informativo*, que no es más que una medida que se asigna a cada concepto de la taxonomía y que mide su nivel de especificación. Un concepto con alto contenido informativo es muy específico, mientras que los conceptos con menos contenido informativo suelen ser de temas menos específicos, más generales. Por ejemplo, “Tenedor” tiene más contenido informativo que “Cosa”. Esto solucionaría los problemas de la no uniformidad en los enlaces y la variedad en la densidad local de los nodos en una jerarquía.

La primera medida que determinó este factor fue la de (Resnik, 1995). La idea es asociar una probabilidad a cada concepto, *p(c).* La probabilidad indica el porcentaje de encontrar una instancia de *c* en base a un corpus, donde se mira la frecuencia de palabras referentes a ese concepto, *frec(c),* que aparecen en ese corpus.

$$p(c) = \frac{frec(c)}{N}$$

Ecuación 6: Fórmula para la probabilidad de (Resnik, 1995)

N es el número total de nombres observados, y *frec(c)* se define de la siguiente manera:

$$frec(c) = \sum_{n \in words(c)} count(n)$$

Ecuación 7: Fórmula para la frecuencia de (Resnik, 1995)

En esta fórmula, *words(c)* es el conjunto de palabras incluidas por el concepto *c*.

Por ejemplo, si sólo hay un concepto en la jerarquía, la probabilidad de encontrar una instancia de ese concepto en el corpus es de 1, luego su contenido informativo será 0.

De esta manera, (Resnik, 1995) calcula la similitud entre dos conceptos c_1 y c_2 de la siguiente manera:

$$sim_{Resnik}(c_1, c_2) = \max_{c \in LCSs(c_1, c_2)} [-\log p(c)] = IC\ (c_1, c_2)$$

Ecuación 8: Medida de similitud de (Resnik, 1995)

LCSs(c_1,c_2) es el conjunto de conceptos que incluyen a c_1 y c_2 en la jerarquía (ancestros).

En la Ilustración 18, se muestra un ejemplo para el cálculo de similitud entre los términos *(Car, Bicycle)* y *(Car, Fork)*. La imagen muestra el fragmento específico que contiene estas clases. El número que se encuentra entre paréntesis indica el valor del contenido de información correspondiente. En la imagen vemos que la similitud entre *Car* y *Bicycle* es el valor del contenido de información de la clase *Vehicle*, que es el valor máximo de las clases que incluyen a ambos (*8,3).* La similitud entre C*ar* y *Fork* es de *(3,53).* Esto confirma que *Car* y *Fork* es menos similar que *Car* y *Bicycles*.

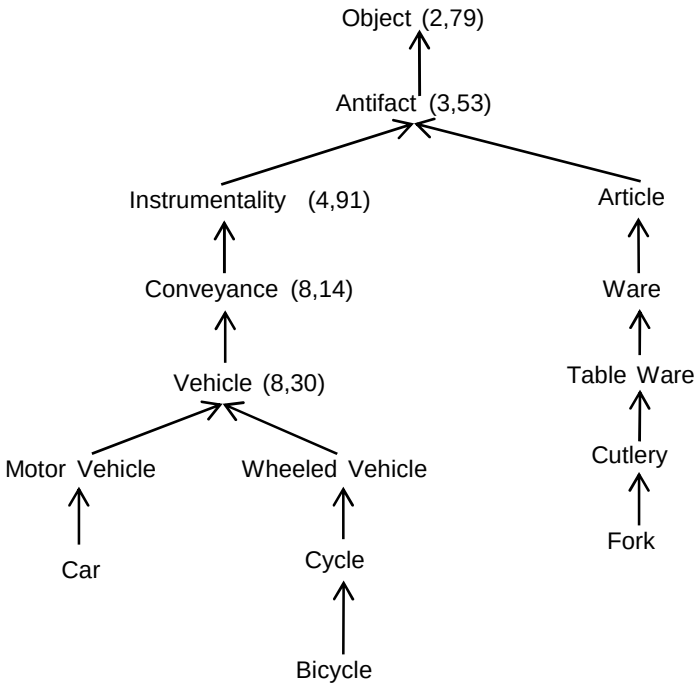

Ilustración 18: Ejemplo ilustrativo del contenido de información en conceptos de WordNet

El problema de la medida de (Resnik, 1995) es que la similitud la hace entre palabras, no entre significados. Esto implica que al buscar las palabras de los conceptos c_1 y c_2 en el corpus, no distingue entre sus posibles significados.

(Richardson & Smeaton, 1995) establecen una medida parecida, pero estableciendo la similitud para significados, no palabras. Cambian el cálculo de la frecuencia *freq(c)*, para tener en cuenta palabras con distintos sentidos, dividiendo la frecuencia utilizada por (Resnik, 1995) por el número de posibles sentidos de la palabra.

$$Frec\ (c) = \sum_{w \in\ words\ (c)} \frac{freq\ (w)}{|\ classes\ (w)|}$$

Ecuación 9: Cálculo de frecuencia de Richardson y Smeaton

Como puede verse, algunos conceptos pueden compartir el mismo *lcs*, por lo que tendrán las mismas medidas de similitud.

Para ello, (Lin & et al., 1998) desarrolla otra versión del trabajo de (Resnik, 1995) proponiendo una medida de similitud parecida, pero normalizando el contenido de información común entre dos conceptos, dado por su *lcs*, con el contenido de información de cada uno de esos dos conceptos individualmente.

$$sim_{lin}(c_1, c_2) = \frac{2\ x\ IC\big(lcs\ (c_1, c_2)\big)}{IC(c_1) + IC\ (c_2)}$$

Ecuación 10: Medida de similitud de (Lin, 1998)

De una manera similar a (Lin & et al., 1998), (Jiang & Conrath, 1997) calculan la similitud semántica, pero escalando la información del *lcs* a través de los conceptos individuales. Sin embargo, este escalado lo hace vía diferencia, en lugar de vía ratio como (Lin & et al., 1998).

$$sim_{Jian}\ (c_1, c_2) = IC\ (c_1) + IC\ (c_2) - 2\ x\ IC\ (lcs\ (c_1, c_2))$$

Ecuación 11: Versión simplificada de la medida de la similitud de (Jiang y Conrath, 1997)

También propone una versión extendida, usando además diferentes factores como la densidad local, la información de contenido, la profundidad del nodo y el tipo de enlace. Utiliza como taxonomía WordNet (en su versión 1.5), y para calcular frecuencias el *SemCor,* un texto etiquetado sacado del *Brown Corpus*.

Otro trabajo importante es el de (Li et al., 2003), que usan una combinación de métricas asumiendo que la similitud semántica depende no sólo de múltiples factores, sino de la correcta combinación entre ellas.

Proponen que la similitud entre dos términos sea una función, considerando 3 atributos:

1) El camino más corto entre dos términos (*length*),
2) la profundidad o altura de un término (*depth*) y
3) el contenido de información.

Recordemos que las dos primeras se consiguen con la taxonomía léxica, y la tercera con un corpus:

$$s(w_1, w_2) = f\ (\ f_1(length), f_2(h), f_3(IC))$$

Ecuación 12: Fórmula lineal para la medida de similitud de (Li et al., 2003)

Cada atributo va entre [0, infinito), pero la similitud debe ir entre [0,1], donde a mayor distancia, menor similitud; es decir, a medida que la distancia decrece a 0, la similitud incrementa monótonamente a 1. Así es que, cada una de las funciones de cada atributo (f_1, f_2 y f_3) se transformarán en funciones no lineales.

Para realizar los cálculos, usan WordNet y Brown Corpus. Teniendo en cuenta las funciones, realizan pruebas para ir probando cómo contribuyen a los resultados cada una de esas funciones.

Las conclusiones tras probar distintas estrategias son:

- En WordNet, la profundidad es un indicador mejor que el camino entre dos nodos.
- Que la información de contenido no influye para nada, así que se elimina de la fórmula final.
- Que el proceso para calcular la similitud no sigue un proceso lineal, sino que es no lineal.

La fórmula final que obtienen es:

$$S(w_1, w_2) = e^{-\alpha \cdot length(w_1, w_2)} \cdot \frac{e^{\beta depth(lcs(w_1, w_2))} - e^{-\beta depth(lcs(w_1, w_2))}}{e^{\beta depth(lcs(w_1, w_2))} + e^{-\beta depth(lcs(w_1, w_2))}}$$

Ecuación 13: Fórmula final de la medida de similitud de (Li et al., 2003)

En nuestro proyecto utilizaremos las medidas que no dependen del contenido de información. Las fórmulas son las siguientes:

$$Li_1(c_1, c_2) = 2 \cdot d - length(c_1, c_2)$$

Ecuación 14: Medida de similitud para el método 1 de (li et al., 2003)

$$Li_2(c_1, c_2) = \alpha \cdot Li_1(c_1, c_2) + \beta \cdot depth\big(lcs(c_1, c_2)\big)$$

Ecuación 15: Medida de similitud para el método 2 de (Li et al., 2003)

$$Li_3 = e^{-\alpha \cdot length}$$

Ecuación 16: Medida de similitud para el método 3 de (Li et al., 2003)

$$Li_4 = e^{-\alpha \cdot length} \cdot \frac{e^{\beta \cdot depth(lcs(c_1, c_2))} - e^{-\beta \cdot depth(lcs(c_1, c_2))}}{e^{\beta \cdot depth(lcs(c_1, c_2))} + e^{-\beta \cdot depth(lcs(c_1, c_2))}}$$

Ecuación 17: Medida de similitud para el método 4 de (Li et al., 2003)

$$Li_5 = \frac{e^{\beta \cdot depth(lcs(c_1, c_2))} - e^{-\beta \cdot depth(lcs(c_1, c_2))}}{e^{\beta \cdot depth(lcs(c_1, c_2))} + e^{-\beta \cdot depth(lcs(c_1, c_2))}}$$

Ecuación 18: Medida de similitud para el método 5 de (Li et al., 2003)

Un problema de estas medidas es la dependencia del corpus. Para determinadas aplicaciones necesitaríamos encontrar corpus adecuados a ellas para obtener buenos resultados.

3.4. Medidas basadas en buscadores Web.

Gracias a la proliferación de los buscadores Web y del aumento de su eficacia y eficiencia, se han podido desarrollar otro tipo de medidas de similitud basada en este tipo de aplicaciones.

La principal ventaja que tiene el uso de buscadores es que casi cualquier palabra posible o significado puede estar indexada, por lo que no se depende de fuentes de datos o vocabularios acotados, donde su descripción podría estar limitada o incluso no existir.

Uno de los primeros trabajos basados en buscadores Web es el desarrollado por (Strube & Ponzetto, 2006). Realiza una medida básica como es tomar los resultados obtenidos al realizar una búsqueda (*hits, page counts*) de un buscador y aplicar el denominado coeficiente de Jaccard.

$$Sim_{Strube}(w_1, w_2) = \frac{Hits\ (w_1\ AND\ w_2)}{Hits\ (w_1) + Hits\ (w_2) - Hits\ (w_1\ AND\ w_2)}$$

Ecuación 19: Coeficiente de Jaccard aplicado a Google para el cálculo de similitud entre dos términos, según (Strube y Ponzetto, 2006)

Un trabajo posterior a este es el de (Cilibrasi & Vitanyi, 2007), que calcula una nueva medida de distancia, NGD (Normalized Google Distance):

$$NGD\ (w_1, w_2) = \frac{G\ (w_1, y) - \min(G(w_1), G(w_2))}{\max(G(w_1), (G(w_2))}$$

$$= \frac{\max\{\log f(w_1), \log f(w_2)\} - \log(f(w_1, w_2)}{\log N - \min\{\log f(w_1), \log f(w_2)\}}$$

Ecuación 20: Medida de distancia de NGD (Cilibrasi y Vitanyi, 2007) donde f(w_x) son los *hits* obtenidos tras la búsqueda de w_x en Google

Otro trabajo que utiliza la medida NGD es el de (Trillo & et al., 2007), que convierte dicha medida de distancia en una medida de similitud, para que quede acotada entre 0 y 1:

$$Google\_rel(w_1, w_2) = e^{-2NGD(w_1,w_2)}$$

Ecuación 21: Transformación de la medida NGD en medida de similitud (Trillo et al., 2007)

Sin embargo, estas medidas miden la relación en general de dos palabras, más que su similitud. Además, (Bollegala & et al., 2007) advierte ciertos problemas en este tipo de medidas, y es que el conteo de resultados ignora la posición de una palabra en una página; es decir, aunque dos palabras estén en la misma página, pueden no estar relacionadas. Aparte, cuentan con el problema de la polisemia, y es que el conteo de resultados de búsqueda para *apple* devuelven las páginas que contienen *apple* como fruta y *apple* como compañía.

Lo que proponen es una medida híbrida que, dada una SVM, combinan 4 medidas basadas en el conteo de resultados de búsqueda y 1 medida basada en *snippets*. Esta nueva medida no mejora a las tradicionales, pero sí mejora las anteriores basadas en buscadores Web, así que su método propuesto puede ser usado para calcular la similitud entre palabras que no están en WordNet u otros tesauros. Las 4 medidas basadas en conteo de resultados de búsqueda son básicamente modificaciones de conocidas medidas de correlación (Jaccard, Dice, etc.). Para la 5ª medida, basada en los snippets, proponen un enfoque basado en la extracción de patrones sintáctico-léxicos. Por ejemplo, en *The jaguar is the largest cat…*, la frase *is the largest* indica una relación de hiperonimia.

3.5. Medidas basadas en Wikipedia

Wikipedia provee una base de conocimiento para calcular la similitud entre palabras de un modo más estructurado que un motor de búsqueda y con más cobertura que vocabularios limitados como WordNet.

Mientras que WordNet representa una taxonomía estructurada, organizada de modo jerárquico, Wikipedia está formada por entradas de un gran número de entidades y conceptos especializados. Wikipedia ofrece una taxonomía pero por medio de sus categorías: los artículos pueden estar asignados a una o más categorías. En la práctica, la taxonomía no está diseñada como una estructura jerárquica, sino que permite múltiples esquemas de categorización que

coexisten simultáneamente, dando lugar a una clasificación de herencia múltiple.

(Strube & Ponzetto, 2006) desarrollan una de los primeros trabajos que usan Wikipedia para medir relaciones de similitud, llamado *WikiRelate!* Aplican medidas de similitud elaboradas para WordNet pero a Wikipedia. Para calcular la similitud entre dos páginas de Wikipedia se valen de 1) una medida de co-ocurrencia (Lesk, 1986), para la cual usan el texto entre las dos páginas; 2) dos medidas basadas en grafos (Leacock & Chodorow, 1994); (Wu & Palmer, 1994)); y 3) una variante de la medida del contenido de información (Resnik, 1995).

Para las medidas basadas en grafos necesitan el *lcs* entre las dos páginas. Para ello, extraen la lista de categorías a las que pertenece cada página. Dadas las listas de categorías, para cada par de categorías, se desarrolla una búsqueda de profundidad máxima de 4 niveles para obtener su *lcs*.

Para calcular el contenido de información no se valen de un corpus aparte, sino del contenido de información intrínseco de un nodo de la propia estructura de categorías, de manera que:

$$IC(cat) = 1 - \frac{log(hipo(cat) + 1)}{\log(C)}$$

Ecuación 22: Contenido de información de una categoría en *Wikirelate!*

En la ecuación, *hipo(cat)* es el número de hipónimos de la categoría *cat,* y *C* el número de nodos en total de la taxonomía.

(Gabrilovich & Marlovich, 2007) elabora una medida que calcula relaciones semánticas entre dos textos arbitrarios, proponiendo un método llamado ESA (*Explicit Semantic Analisys*). Esta técnica pertenece a las medidas basadas en co-ocurrencia y representa cada concepto de Wikipedia como un vector de palabras que ocurren en el artículo del concepto, con un determinado peso dado por la técnica de *tf/idf*. Construyen entonces un índice invertido donde, a cada palabra se le asigna la lista de conceptos de Wikipedia donde aparece. Finalmente, para calcular la relación semántica entre dos textos, representan cada texto como un vector de palabras y, por cada una de esas palabras, recuperan del índice invertido la lista de conceptos de Wikipedia donde aparecen, obteniéndose así dos vectores multidimensionales sobre los que se aplicará la medida del coseno.

Otro trabajo parecido es el de (Hassan & Wee, 2008), donde miden la similitud de sus palabras con el coseno de sus vectores, que también son una lista de conceptos de Wikipedia donde aparecen.

También podemos encontrar otros trabajos como *WLM* (*Wikipedia Link-based Measure*) de (Milne & Witten, 2008), que calculan relaciones en función de los hiperenlaces que se encuentran dentro de los artículos de Wikipedia. Obtienen la medida como combinación de dos métricas (realizando una media entre ellas). La primera la obtienen a partir del ángulo entre los vectores de los enlaces encontrados dentro de los dos artículos. Es similar a la técnica *tf/idf*, pero en lugar de trabajar con los pesos obtenidos con la probabilidad de encontrar cada término, trabaja con la probabilidad de encontrar cada enlace.

Por lo tanto, si c_1 y c_2 son los conceptos origen y destino, entonces el peso *w* del enlace $c_1 \rightarrow c_2$ es:

$$w(c_1 \rightarrow c_2) = log\left(\frac{|W|}{|C_2|}\right)$$

Ecuación 23: Primera métrica para la medida WLM

Donde *W* es el conjunto de todos los conceptos de Wikipedia y C_2 el número de conceptos que enlazan con c_2. Por lo tanto, los enlaces son considerados menos significativos para juzgar la similitud entre artículos si otros muchos artículos son enlazados al mismo c_2. Estos pesos obtenidos para cada enlace serán usados para generar vectores que describen cada uno de los dos conceptos en cuestión. Para terminar se calcula el coseno.

La segunda métrica de *WLM* es parecida a la medida *NGD* pero, en lugar de trabajar con los resultados de búsquedas, trabaja con enlaces de Wikipedia:

$$sim(c_1,c_2) = \frac{\log(\max(|C_1|,|C_2|)) - log(|C_1 \cap C_2|)}{log(|W|) - log(\min(|C_1|,|C_2|))}$$

Ecuación 24: Segunda métrica para WLM

C_1 y C_2 son los conjuntos de artículos que enlazan a c_1 y c_2 respectivamente. $C_1 \cap C_2$ representa el factor de co-ocurrencias, y se refiere a las páginas de Wikipedia que enlazan a ambos conceptos.

Otro trabajo que utiliza los enlaces de Wikipedia para calcular relaciones semánticas es el de (Zhang et al., 2011). Hace una diferenciación entre relaciones implícitas y explícitas. Una relación explícita viene dada por un

hiperenlace entre dos conceptos. Una implícita viene dada por una estructura de múltiples enlaces y páginas entre los dos conceptos.

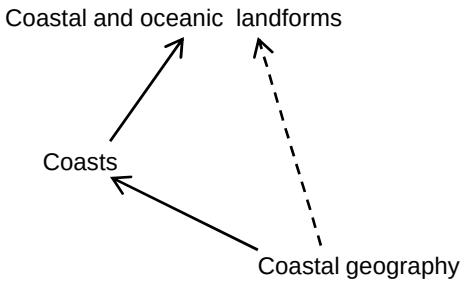

Ilustración 19: Ejemplo de relación explícita (Coastal geography -> Coasts) e implícita (Coastal geography -> Coastal and oceanic landforms, a través de Coasts)

Así, calculan la fuerza de la relación semántica desde c_1 hasta c_2 usando el valor del flujo entre c_1 y c_2. Cada arco tiene un peso. El flujo enviado a lo largo de cada arco es multiplicado por un peso, que es asignado a cada arco a través de una función utilizando la estructura de Wikipedia.

Esta métrica usa tres factores: la distancia, la conectividad y la co-citación. La distancia es la longitud del camino más corto entre conceptos, como ya hemos visto en otros trabajos. La conectividad desde c_1 hasta c_2, es el número mínimo de vértices necesarios para que puedan existir caminos entre c_1 y c_2 (cuanta más conectividad, más relación). Finalmente, co-citación es el inverso de la co-ocurrencia, y mide el número de enlaces a ambos conceptos (relación más fuerte cuanto más grande sea el número).

El principal problema de todas estas medidas es que miden la relación general entre dos páginas, y no su similitud semántica en concreto. Además, los resultados obtenidos con dichas medidas no consiguen acercarse a las medidas de similitud tradicionales, como veremos a continuación en el siguiente capítulo.

4. ANÁLISIS DEL PROBLEMA

Una vez se han presentado y expuesto el contexto del problema junto con algunas de las soluciones aportadas hasta el momento, se pasa a analizar los problemas encontrados en dichas soluciones.

En este capítulo además se analizan los objetivos del proyecto junto las principales características de nuestra solución basada en Wikipedia.

4.1. Planteamiento del problema

Como se ha explicado en el capítulo “3: Estado del arte”, podemos encontrar muchas medidas que calculan la similitud semántica entre dos términos. Sin embargo, esos enfoques tienen que lidiar con algunos problemas, la mayoría relacionados con la fuente de datos sobre la cual operan. Gran parte de las medidas toman taxonomías como WordNet, diccionarios o colecciones grandes de documentos como su fuente principal de información, con claras desventajas, que pueden resumirse en que:

1) Si usan fuentes demasiado generales, no pueden utilizarse para términos específicos, nombres propios, etc., o que requieran cubrir un amplio número de conceptos de la vida real.

2) Si funcionan bien con determinados corpus de un área concreta, dejan de ser independientes del dominio.

3) Suelen apoyarse de fuentes de datos donde cualquier término nuevo o modificación tarda bastante en ver la luz.

Existen otras medidas que usan la propia Web como un gran corpus, usando los motores de búsquedas para calcular la similitud. Esto garantiza que casi cualquier término que se busque se va a encontrar, pero ya no se pueden valer de factores basados en jerarquías y estructuras taxonómicas en general.

Wikipedia, no obstante, ofrece por un lado un conocimiento mucho más amplio que taxonomías existentes y, por otro lado, está estructurada de una manera más definida que la información que puede existir en la Web. El conocimiento que alberga Wikipedia está consensuado por un gran número de personas y se actualiza de manera ágil, además de estar traducida en numerosas lenguas. Es por esto que una medida de cálculo de similitud semántica basada en esta fuente parece resultar, en principio, la más idónea y eficaz, cubriendo no sólo

multitud de dominios y escenarios sino que también puede utilizarse para procesar términos de diferentes idiomas.

4.2. Objetivos del proyecto

Como principio básico para solventar los problemas citados anteriormente, asumiremos que *es posible el uso eficaz de la Wikipedia como fuente de datos para calcular la similitud semántica entre dos conceptos*. Por esto, se fijarán dos objetivos:

- Descargar y almacenar Wikipedia en una base de datos local, guardando los conceptos, categorías y las relaciones entre ellas.
- Conseguir iguales o mejores resultados aplicando la Wikipedia a medidas tradicionales que anteriormente usaban otras taxonomías.

Para la consecución de estos dos objetivos se tendrá en cuenta un único requisito, y es que el uso de Wikipedia debe ser tal que se evite el procesamiento de ingentes cantidades de documentos y el uso de posibles técnicas tediosas de procesamiento de lenguaje natural. Para ello, dicho procesamiento hará uso únicamente de la estructura de categorías en las que están clasificados los artículos de Wikipedia, sin tener en cuenta el contenido de información de los artículos en sí. Esto no impide que en un futuro pueda utilizarse para mejorar los resultados que se obtengan una vez finalizado este proyecto.

4.2.1. Almacenamiento de Wikipedia

Para el desarrollo de los objetivos propuestos, será necesario guardar la estructura de categorías de Wikipedia. Al almacenar esta estructura habrá que tener en cuenta su peculiar organización, y es que el sistema de categorización de artículos de Wikipedia cuenta con dos características que la alejan de ser considerada una taxonomía convencional:

- Por su estructura: No sigue una estructura jerárquica propiamente dicha (por ejemplo, podemos encontrarnos con caminos cíclicos).
- Por su utilización a la hora de categorizar conceptos: El sistema de categorización de conceptos se asemeja más a un sistema de etiquetado (*tagging system*) que a un sistema jerárquico con relaciones de hiperonimia como WordNet.

En este proyecto se trabajará sólo con la versión inglesa de Wikipedia.

4.2.2. Aplicación de Wikipedia a medidas tradicionales

Procesando adecuadamente los factores tradicionales de jerarquías (longitud entre dos nodos, profundidad o altura de un nodo, etc) en la información de Wikipedia almacenada, se procederá a implementar las medidas tradicionales más relevantes basadas en caminos. No se implementarán medidas basadas en el contenido de información, ni en técnicas basadas en corpus, al no formar parte de nuestro principio básico (recordemos que trabajamos sólo con una estructura de categorización y no con técnicas de procesado de textos).

El motivo de dividir el conjunto original de pares de R&G es para asegurarnos de que la medida final no se haya "viciada" por un determinado conjunto concreto. Si realizamos una fórmula "a medida" para el conjunto de prueba, como se hace por ejemplo en (Blázquez-del-Toro et al., 2008), lógicamente maximizaremos los resultados obtenidos, pero no nos asegurará que funciona para otro conjunto de datos.

4.3. Conjunto de datos para los experimentos

La calidad de un método computacional para calcular la similitud semántica se puede establecer comparando su resultado con muestras de datos obtenidas de personas físicas.

Una muestra de datos que suele utilizarse para el cálculo de similitud semántica es la de (Rubenstein & Goodenough, 1965) a partir de ahora R&G, que publicaron su trabajo sobre similitud entre términos en contextos relacionados. Para testar sus afirmaciones, seleccionaron 65 pares de nombres cuya similitud fue evaluada por dos grupos de personas de 15 y 36 sujetos, respectivamente. Esta similitud se calificó entre 0.0 (no similares) y 4.0 (sinónimos).

Este conjunto de términos y sus valores de similitud son considerados un estándar dentro este campo. Así pues, las medidas de similitud sólo tienen que ver cómo correlacionan sus resultados con aquellos de R&G para saber su eficacia. Por este motivo, utilizaremos el conjunto de palabras del experimento de R&G y sus valores. Este conjunto se dividirá en dos subconjuntos: un primero de 28 pares, que llamaremos *conjunto de prueba*, y un segundo conjunto de 37 pares que llamaremos *conjunto de entrenamiento*. El subconjunto de prueba fue seleccionado al azar por primera vez en el trabajo de (Miller & Charles, 1991), que replicaron el experimento de R&G con 38

sujetos. El segundo subconjunto de 37 pares lo utilizaremos como conjunto de entrenamiento para la obtención de factores adecuados en nuestros cálculos.

4.4. Resultados publicados de medidas tradicionales

La Tabla 10 muestra los coeficientes de correlación de Pearson para el conjunto de prueba, de las medidas de similitud más relevantes vistas en el capítulo anterior.

Tabla 10: Coeficientes de correlación para el conjunto de prueba de las medidas existentes más relevantes

Método	Tipo	Correlacion publicada
Rada et al., 1989 (camino más corto)	Grafo	0.66
Wu y Palmer, 1994	Grafo	0.79
Leacock y Chorodow, 1994	Grafo	0.83
Blazquez-del-Toro et al., 2008	Grafo	0.81
Resnik, 1995	Múltiples fuentes	0.74
Lin, 1998	Múltiples fuentes	0.75
Jiang y Conrath, 1997	Múltiples fuentes	0.84
Li et al., 2003	Múltiples fuentes	0.89
Bollegala et al., 2007	Buscadores Web	0.79
WikiRelate!, 2006	Wikipedia	0.56
Gabrilovich y Markovitch, 2007	Wikipedia	0.75
Wee y Hassan, 2008 (*)	Wikipedia	0.60
Milne y Witten, 2008	Wikipedia	0.64
Zhang et al., 2010 (*)	Wikipedia	0.56

(*) No usan el conjunto estándar de R&G para valorar sus experimentos, así que la correlación publicada no es en base a ese conjunto

Los coeficientes más altos (mayores de 0.8) se han obtenido a partir de medidas basadas en múltiples fuentes y medidas basadas en grafos. Sólo una de ellas (Jiang & Conrath, 1997) usa el contenido de la información como factor básico.

El método de (Bollegala & et al., 2007) cubre de una manera más amplia cualquier dominio, al contar con un motor de búsqueda web como fuente, pero sus resultados no son mejores que los obtenidos por medidas basadas en grafos, como la de (Wu & Palmer, 1994).

Los modelos que usan Wikipedia tampoco ofrecen resultados prometedores. La mejor métrica de este grupo es la de WikiRelate!, que usa la medida del camino más corto. La medida de (Milne & Witten, 2008) está basada en los hiperenlaces que hay entre artículos, así que es de bajo coste computacional,

pero sus resultados no se acercan a los de (Gabrilovich & Marlovich, 2007) o (Bollegala & et al., 2007).

4.5. Estructura de categorías de Wikipedia

El esquema de categorización que sigue Wikipedia es el de un grafo dirigido cíclico; no sigue una estructura jerárquica acíclica bien definida, como es el caso de WordNet, aspecto que habrá que tener en cuenta para su manejo.

4.5.1. Estableciendo una raíz

Un ejemplo de los primeros niveles de la estructura de categorías lo vemos en la Ilustración 20. La raíz de la jerarquía es la categoría *Cat: Contents*. A partir de la raíz podemos tener múltiples caminos o *ramas* para llegar a cada uno de los nodos. El último elemento de cada camino, aquel que no tiene hijos, se denomina nodo *hoja*.

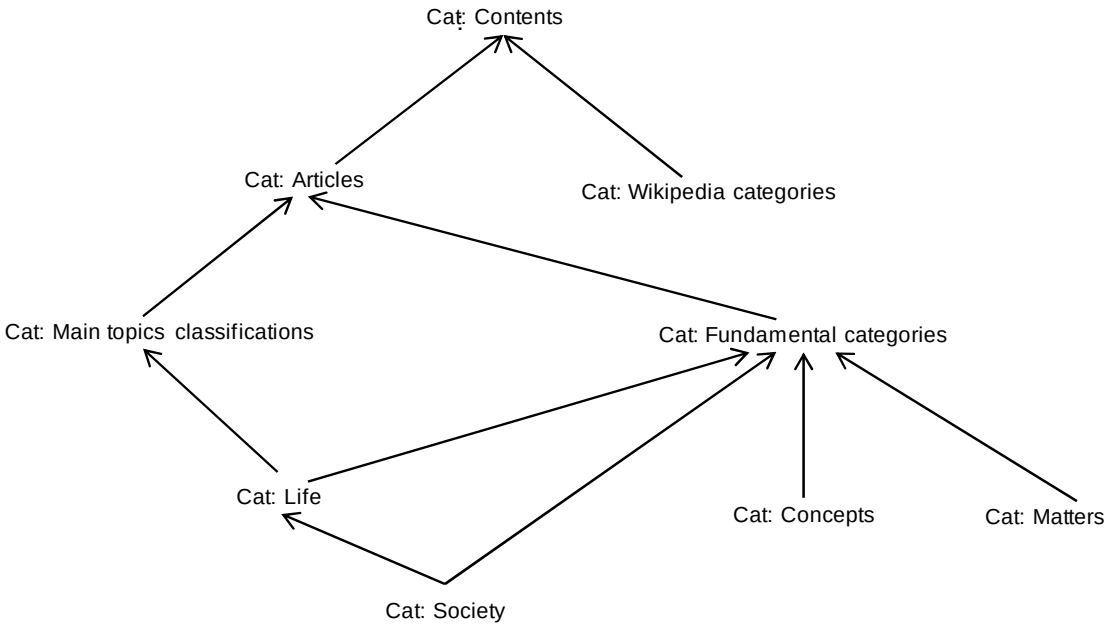

Ilustración 20: Primeros niveles de la jerarquía de categorías en Wikipedia

La categoría raíz agrupa todos los diferentes tipos de páginas existentes en Wikipedia. Sin embargo, para llevar a cabo nuestro cometido, sólo nos interesa la subjerarquía de la categoría *Cat: Articles*, que es la categoría que agrupa a los artículos por contenido. Las categorías que se encuentran fuera de este subárbol suelen ser de otro tipo no ligado a este proyecto; por ejemplo, suelen agrupar páginas administrativas, o clasificar los artículos por estado. De entre

sus categorías hijas principales, es *Cat: Fundamental categories* la que clasifica todos los artículos de una manera más lógica y progresiva, así que esta categoría forma, junto con sus subcategorías, la fuente de datos con la que se trabaja durante todo el proyecto.

4.5.2. Gestión de ciclos

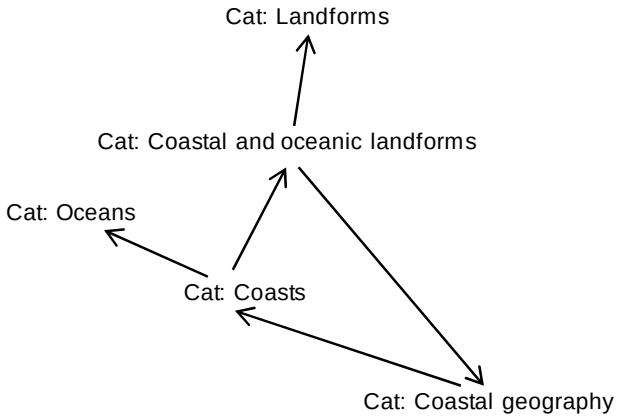

Ilustración 21: Ejemplo de grafo dirigido cíclico en Wikipedia

En el ejemplo de la Ilustración 21, se muestra un ejemplo de ciclo, en el que la categoría *Cat: Coastal and oceanic landforms* está incluida en la categoría *Cat: Coastal geography*, que a su vez está incluida en la categoría *Cat: Coasts*, y que a su vez se incluye en la categoría *Cat: Coastal and oceanic landforms*. Durante el procesamiento de la estructura de Wikipedia, hay que tener en cuenta dichos ciclos para poder procesar los caminos sin bucles infinitos.

4.5.3. Herencia múltiple

Otra característica importante de la estructura de Wikipedia es la herencia múltiple entre categorías y entre categorías y conceptos.

La herencia múltiple entre categorías se puede ver en la Ilustración 22, donde *Cat: Fruit* tiene como categorías padre directas a 3 categorías más.

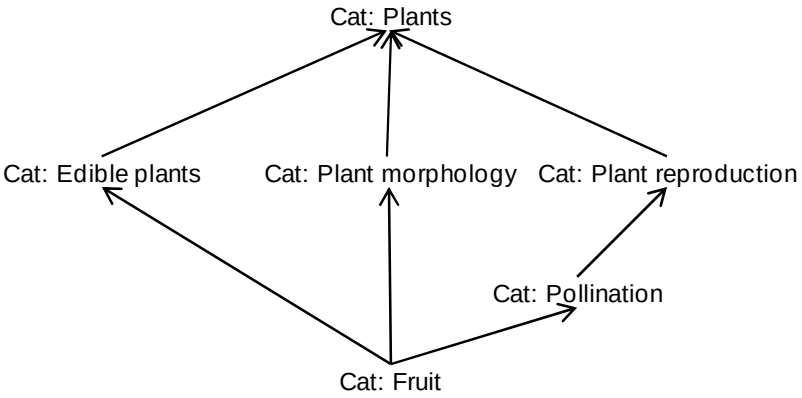

Ilustración 22: Ejemplo de herencia múltiple en Wikipedia

Respecto a la herencia múltiple entre categorías y conceptos, la clasificación de Wikipedia podría considerarse un sistema de etiquetado (*tagging system)* más que una clasificación propiamente dicha. Un ejemplo muy representativo de esto es el artículo referido a *Barack Obama*. En la Ilustración 23, podemos comprobar las categorías a las que pertenece el artículo citado. Se ve que más que una categorización de la página es un etiquetado de la misma.

Ilustración 23: Ejemplo de categorización de Wikipedia para uno de sus artículos

4.6. Modelo Conceptual

Wikipedia cuenta con gran cantidad de información de la que sólo una porción será relevante para la consecución de nuestro proyecto; básicamente interesa su estructura de categorías y los propios artículos de los conceptos. Así es que,

antes de pasar a describir el diseño del proyecto, se detalla aquí el modelo conceptual básico (modelo de información estática), que representa las entidades que se van a almacenar en la aplicación para su posterior procesamiento. Para ello se utiliza la técnica UML de los *diagramas de clases*. Este tipo de diagramas presenta las clases de la aplicación con sus relaciones estructurales y de herencia.

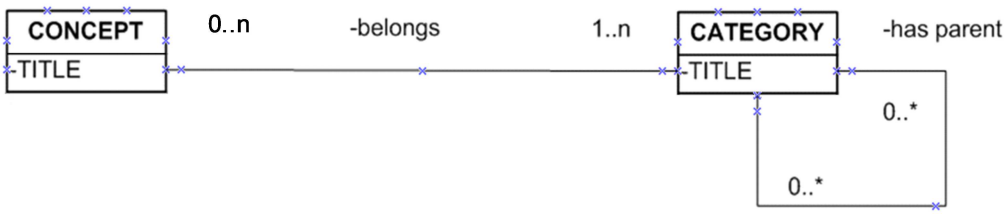

Ilustración 24: Modelo conceptual

La Ilustración 24 muestra un modelo bastante sencillo. Las clases son:

- *Concept*: Representa a una página o artículo de Wikipedia con información de una entidad concreta del mundo real. Su URL es:

  http://en.wikipedia.org/wiki/[title]

  Por ejemplo:

  http://en.wikipedia.org/wiki/Shore

En este tipo de artículos no entran 1) páginas de desambiguación; 2) páginas de redirección; y 3) páginas de listas. La similitud semántica se calculará entre estos conceptos. De aquí en adelante y a lo largo de todo el documento, identificaremos a estos conceptos con su nombre local (atributo *title* en la Ilustración 24), sin el espacio de nombres; por ejemplo, *Shore*.

- *Category:* Los conceptos se engloban dentro de categorías. Dicha categorización será la que nos permitirá obtener las relaciones entre los diferentes pares de conceptos. Su URL es:

  http://en.wikipedia.org/wiki/Category:[title]

  Por ejemplo:

  http://en.wikipedia.org/wiki/Category:Coasts

De aquí en adelante y a lo largo de todo el documento, identificaremos a las categorías con su nombre local (atributo *title* en la Ilustración 24), sin el espacio de nombres, y precedidos del prefijo *Cat:* por ejemplo, *Cat: Coasts*.

No se tiene en cuenta la siguiente información para las clases:

- Contenido de los conceptos: Desarrollaremos el proyecto en base al sistema de categorías de Wikipedia. El contenido de texto de cada artículo podrá servir en un futuro.
- Idioma: Este proyecto trabajará, como hemos comentado anteriormente, sólo con la versión en inglés por comodidad, así que no será necesario especificar el idioma de los conceptos.
- URL: La URL se puede deducir del título, ya que ésta estará formada del sufijo *http://en.wikipedia.org/wiki/Category:* más el atributo *title* de la categoría, o bien el sufijo "http://en.wikipedia.org/wiki/" más el atributo *title* del concepto.

Entre las relaciones que se tendrán en cuenta, se encuentran:

- *belongs to*: Un concepto pertenece (al menos) a una categoría. Mientras, una categoría puede contener muchos conceptos o ninguno.
- *has parent*: Las categorías se conectan unas con otras dentro de relaciones jerárquicas de hiperonimia. Una categoría puede no tener ninguna categoría hija (el ya comentado nodo hoja*)* o muchas. A su vez, una categoría puede tener muchas categorías padres o ninguna; en este último caso, será la categoría raíz.

5. DISEÑO GENERAL

Este capítulo, una vez analizado el sistema, expone las decisiones tomadas a la hora del diseño general de la aplicación, así como las explicaciones de cada uno de los componentes utilizados en la arquitectura del diseño. Se exponen así los componentes del sistema mostrando su arquitectura, parte básica para la definición detallada del diseño que se hará en el siguiente capítulo, descomponiendo el sistema tanto en sus elementos software como en sus componentes hardware.

5.1. Componentes del sistema

Este proyecto, no requiere una arquitectura excesivamente compleja para su desarrollo porque además no cuenta con una interfaz gráfica. Sin embargo, y con vistas a posibles extensiones futuras, se descompondrá en componentes claramente diferenciados. En la Ilustración 25 se puede ver la arquitectura lógica del proyecto, que ha seguido una de las arquitecturas típicas basada en capas (Larman, 2005).

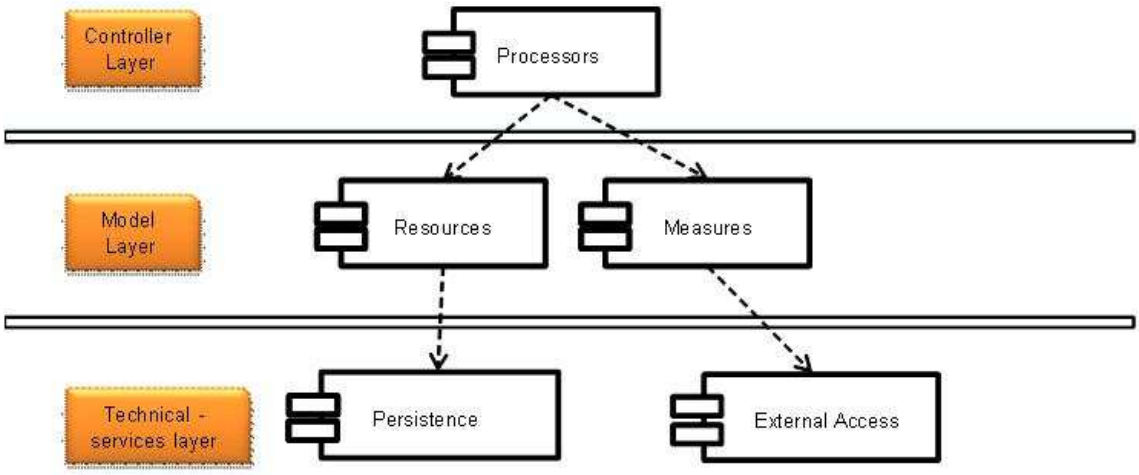

Ilustración 25: Diagrama de componentes

5.1.1. Capa controladora

La capa controladora (*Controller layer*), tiene un controlador principal, *Processors*. Más tarde en el diseño detallado se verá como este componente principal tiene dos componentes subordinados, *CrawlingProcessor* y *SimilarityProcess.*

5.1.2. Capa de modelo

Dentro de esta capa se almacenan tanto los objetos que conforman el dominio de la aplicación como la lógica de negocio y reglas que lo gestionan.

El componente *Resources* se encarga de los diferentes objetos de la estructura de Wikipedia y sus relaciones entre ellos (categorías, parejas de conceptos, etc.). El componente *Measures* se encarga de almacenar los procedimientos y funciones necesarios para el cálculo de las diferentes medidas.

5.1.3. Capa de los servicios técnicos

El servicio técnico fundamental es el de la persistencia (componente *Persistencia*) de los elementos modelados en la capa de modelo anterior (componente *Resources*).

6. DISEÑO DETALLADO

Completada la fase de diseño general, en este capítulo se menciona las herramientas y lenguajes que se han utilizado para la implementación. Así mismo, para que dicha implementación y futuros mantenimientos se lleve a cabo, se especifica en detalle los distintos componentes, de tal manera que se vea su interacción para cumplir los objetivos del proyecto.

6.1. Elección de herramientas y lenguajes

Para la realización de la implementación, y por tanto, del diseño detallado, se necesita conocer cuáles van a ser las herramientas y el lenguaje a utilizar.

El lenguaje utilizado para implementar el proyecto es Ruby, en su versión 1.8. El gestor de base de datos donde se almacena la información de Wikipedia, y sobre la que se procesará su estructura, es MySQL Server 5.1.60.

Todo ello se implementará sobre un sistema operativo de software libre como es Fedora 14[2], aunque podría haberse realizado sobre cualquier otro sistema operativo de entorno Windows, Linux, Macintosh.

6.1.1. Ruby

Ruby[3] es un lenguaje de *script* interpretado para una programación orientada a objetos rápida y fácil. Su elección se debe a que es un lenguaje simple, directo, extensible, gratuito y portable. Las características básicas con las que cuenta Ruby son:

- Potencia: Combina la energía pura de la orientación a objetos con la expresividad que puede tener un lenguaje interpretado. Los programas son compactos a la vez que legibles y fáciles de mantener.
- Simplicidad: Cuenta con una sintaxis intuitiva y limpia. Ruby no necesita declaraciones de variables y usa una convención de nombrado sencilla para denotar el alcance de las variables.
- Disponibilidad: Ruby es de código libre y gratuito.
- Portabilidad: La elección de la plataforma ya no es un problema porque un programa en Ruby puede ejecutarse de manera indistinta en Windows, Solaris, Linux, Macintosh, etc.

[2] Página oficial de Fedora versión 14: http://fedoraproject.org/wiki/Releases/14/Schedule
[3] Página oficial del lenguaje de programación Ruby: http://www.ruby-lang.org/es/

6.1.2. MySQL Server

MySQL Server[4] es uno de los servidores más populares de bases de datos relacionales, desarrollado y proporcionado por *MySQL AB*. MySQL Server es un sistema de administración de bases de datos relacionales; almacena los datos en tablas separadas en lugar de en un solo lugar, agregando flexibilidad y velocidad. Las tablas se enlazan al definir relaciones que hacen posible combinar datos de varias tablas cuando se necesita consultar datos. MySQL Server es de código abierto, por lo que se puede usar y modificar libremente.

Los principales motivos para su elección son su rapidez, su seguridad, y su facilidad de uso.

6.2. Diseño de esquema de la base de datos relacional

El diseño lógico de la información necesaria viene representado en la Ilustración 26. Toda la estructura de Wikipedia será almacenada en las tablas y relaciones siguientes:

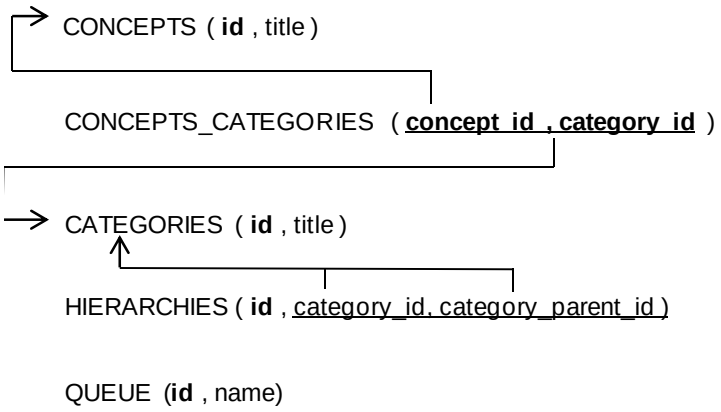

Ilustración 26: Esquema lógico de la base de datos relacional

Las columnas que son clave primaria se encuentran en negrita. Las columnas que son clave ajena se encuentran subrayadas.

Excepto la tabla *Concepts_Categories*, todas las tablas tienen el campo *id* como identificador del registro de esa tabla, que será la clave primaria de la misma.

[4] Página de descarga del servidor MySQL, http://dev.mysql.com/downloads/mysql/

La descripción de las cinco tablas es la siguiente:

- Concepts: Representa la clase Concept de la Ilustración 24. Almacena cada uno de los artículos de Wikipedia. La única información a almacenar de cada artículo va a ser el título del mismo, almacenado en el campo title.
- *Categories*: Representa la clase *Category* de la Ilustración 24. Almacena cada una de las categorías de Wikipedia. El nombre de la categoría será lo que quedará almacenado en el campo *title*. Recordemos que se almacenarán las categorías *Cat: Fundamental categories*, y sus hijas directas e indirectas.
- *Concepts_Categories*: Representa la relación *belongs to*, y queda representada como una tabla intermedia ya que no tiene entidad propia.
- *Hierarchies*: Representa la relación *has parent* entre categorías. Almacenamos cada relación padre-hijo. Si una categoría cat_m contiene tres subcategorías $subcat_x$, $subcat_y$, $subcat_z$, se almacenan tres registros:
  - (id_n, $subcat_x$, cat_m)
  - (id_{n+1}, $subcat_y$, cat_m)
  - (id_{n+2}, $subcat_z$, cat_m)
- *Queue*: Tabla a utilizar durante el procedimiento de descarga y almacenamiento de categorías (a diferencia del resto de tablas que se utilizarán en el procesado de la estructura de Wikipedia). En esta tabla se encolan las categorías que tienen que ser procesadas para su posterior almacenamiento en base de datos. La tabla se compone de un identificador *id* y del nombre de la categoría en cuestión *name*.

6.3. Conjuntos de conceptos a tratar

Como ya se comentó en el apartado "4.3 Conjunto de datos para los experimentos", vamos a considerar dos conjuntos de pares de palabras, a partir de los 65 utilizados por el conjunto original del experimento de R&G:

- Conjunto de entrenamiento (conocido como *training set*): Está formado por 37 pares de conceptos y se utilizarán para las diversas pruebas dentro de la propia realización del proyecto.
- Conjunto de prueba (conocido como *evaluation set)*: Está formado por 28 pares y se utilizará para evaluar los resultados conseguidos una vez fijadas las diferentes fórmulas de cálculo de similitud semántica y poder compararse con trabajos anteriores.

Estos 65 pares de palabras son términos que no se encuentran desambiguados, por lo que en primer lugar habrá que desambiguarlos, obteniendo el artículo de Wikipedia correspondiente al concepto en cuestión.

Al buscar una palabra en Wikipedia, podemos encontrarnos con los siguientes casos:

- La palabra guarda una relación de sinonimia con otras que hacen referencia al mismo significado. Wikipedia en estos casos puede no tener una página de contenido para cada palabra y así evitar redundancia, teniendo sólo una página para ese significado (concepto). Si la palabra en cuestión no corresponde literalmente con la página de ese concepto, será redireccionado a ésta. En la Ilustración 27 se puede ver un ejemplo, donde los términos *Car, Cars* y *Automobile* están asociados a un único concepto en Wikipedia, *Automobile.*

Ilustración 27: Ejemplo de redirección en Wikipedia

- La palabra es polisémica y tiene varios significados, luego puede estar asociada a varios conceptos. Como Wikipedia no sabe a qué significado de la palabra nos referimos, presenta una página de desambiguación con enlaces a cada posible concepto. Las páginas de desambiguación en Wikipedia aparecen bajo la categoría de *Disambiguation pages*, pudiendo contener el sufijo *_disambiguation* (versión inglesa) detrás del nombre a desambiguar, aunque también puede haber páginas de desambiguación que no presenten dicho sufijo.

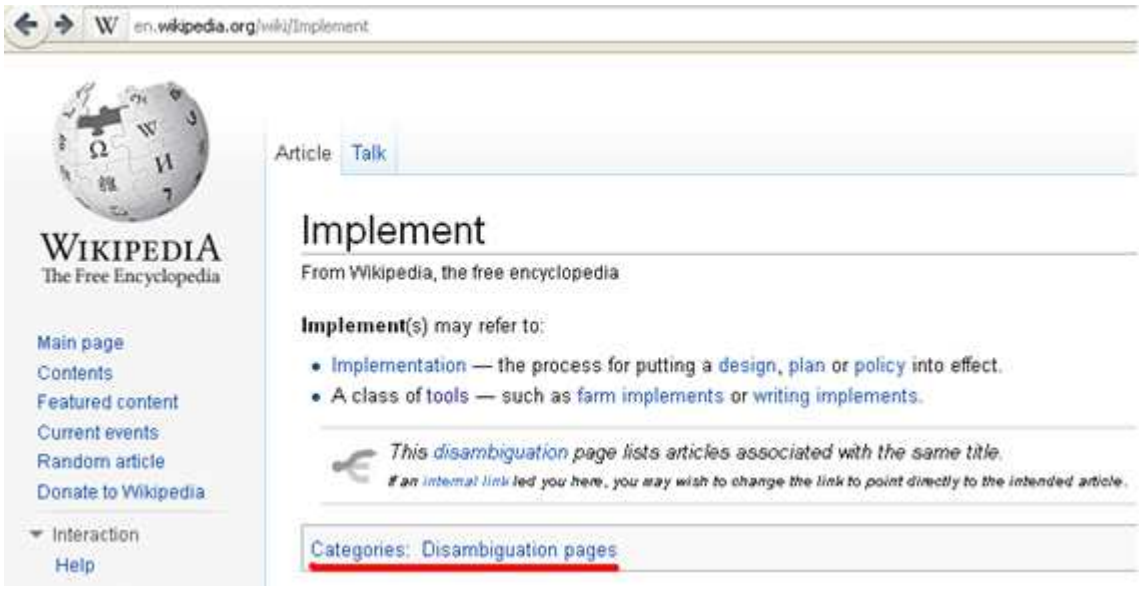

Ilustración 28: Ejemplo de página de desambiguación en Wikipedia

- Una palabra puede hacer referencia a un término muy general para el que Wikipedia tiene una página, pero esa página no representa un concepto específico, sino una lista de conceptos.

Las palabras a desambiguar serán todas aquellas que no dispongan de una página para el concepto correspondiente (artículo de información propiamente dicho, sin ser éste una página de desambiguación, una lista, o una página de redirección a otra). Si podemos encontrar el artículo de Wikipedia de una palabra, esa palabra estará desambiguada. Un modo práctico para saber si el concepto está desambiguado será buscar dentro de nuestra base de datos, ya que almacena todos los conceptos de Wikipedia. Dentro de esta base de datos sólo se encuentran conceptos propiamente dichos; no se almacenan páginas de desambiguación, listas o redirecciones.

Como no hay posibilidad de saber a qué concepto exacto se referían en cada palabra del conjunto de pares del experimento de R&G, se ha elegido finalmente, en el caso de palabras ambiguas, el concepto que representa a un sustantivo y que ha ofrecido mejores resultados. La lista de esos conceptos seleccionados puede verse en el Apéndice A.

6.4. Diseño detallado

Aquí se recogerá de una manera más detallada el diseño de la aplicación, dividida en los componentes que se vieron en el apartado “5.1. Componentes del sistema”, del capítulo anterior.

6.4.1. Capa controladora

La capa controladora está formada por un solo componente, *Processors*, formada a su vez por dos clases principales, que realizan dos funcionalidades distintas y totalmente independientes (ver Ilustración 31; los elementos que aparecen en gris pertenecen a otras capas, definidas en siguientes apartados).

Por un lado tenemos la clase *CrawlingProcessor*, que se encargará del proceso relacionado con la descarga y almacenamiento de Wikipedia en la base de datos. Dicha clase irá procesando cada una de las categorías de Wikipedia, a partir de la categoría que hemos considerado raíz (*Cat: Fundamental categories*). El proceso será iterativo, y se repetirá para cada categoría de Wikipedia hasta recorrer toda su estructura por completo. El proceso efectuado para cada categoría puede resumirse en los siguientes 3 puntos:

1) Para procesar cada categoría de Wikipedia, el primer paso es obtener el código fuente de cada categoría en cuestión. En el ejemplo de la Ilustración 29, se muestra la página de la categoría *Cat: Nature_deities*.

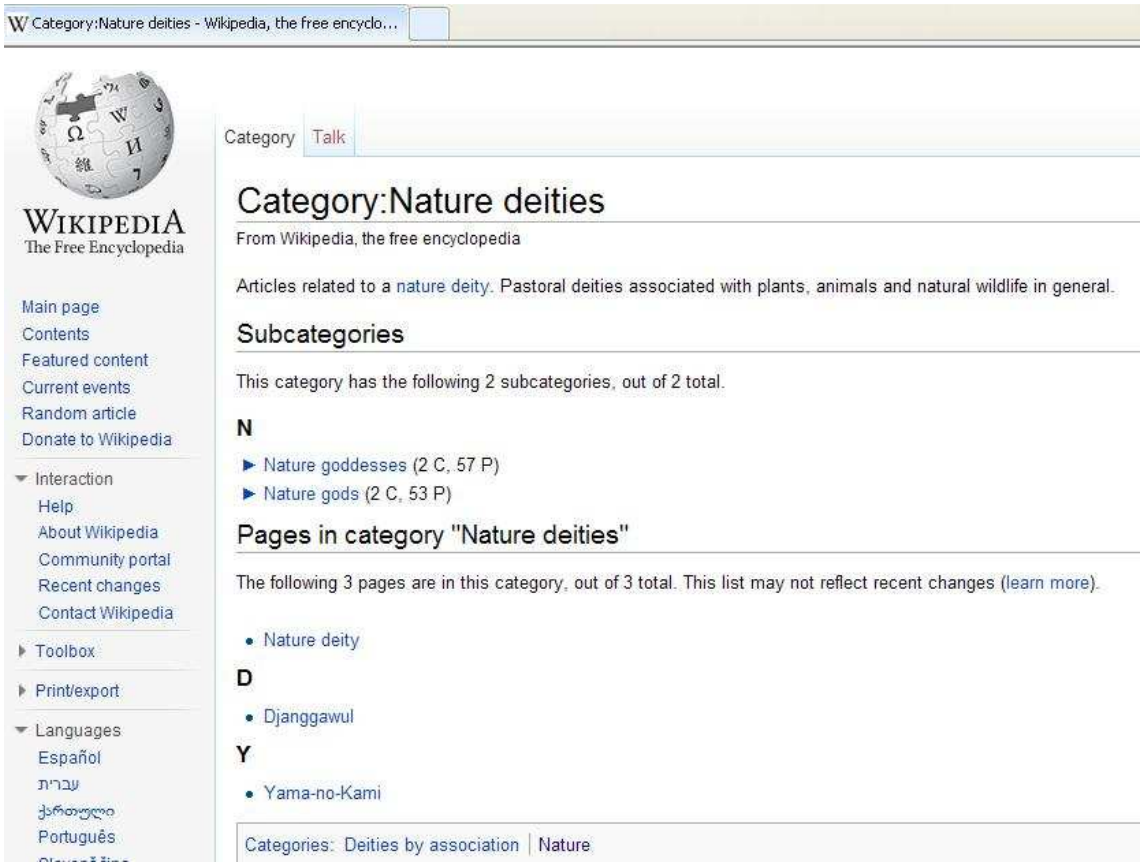

Ilustración 29: Ejemplo de página de una categoría de Wikipedia

2) Mediante el código fuente, se obtienen las subcategorías que tiene dicha categoría y los conceptos que posee. En el ejemplo de la Ilustración 29, vemos que *Cat: Nature deities*, posee 2 subcategorías y 3 conceptos (grupo *Pages*).
3) Se almacenan en la base de datos los siguientes registros:
    a. La categoría que estamos procesando (si es que todavía no existe en la base de datos, en la tabla *Categories*). En el ejemplo de la Ilustración 30 (zona y flecha a), se almacenaría la categoría *Cat: Nature deites.*
    b. Los conceptos que pertenecen a la categoría en la tabla *Concepts*. En el ejemplo de la Ilustración 30 (zona y flecha b), se almacenaría un registro para el concepto *Nature_deity*, otro para *Djanggawul,* y otro para el concepto *Yama-no-Kami.*
    c. La relación de pertenencia de los conceptos a la categoría procesada en la tabla *Concepts_Categories*. En el ejemplo anterior, se almacenarían tres registros dentro de la tabla *Concepts_Categories*, uno con el identificador de la categoría *Cat: Nature_deities* (*692*) y el identificador del primer concepto (*26491*), otro registro igual pero con el identificador del segundo

concepto (*26492*) y un tercero con el identificador del tercer concepto (*26493*) (ver Ilustración 30, tabla c).

d. La relación jerárquica entre la categoría que estamos procesando y su categoría padre (que habría sido almacenada previamente). Será almacenada en la tabla *Hierarchies* (ver Ilustración 30, zona y flecha d).
e. Por último, se almacena en la tabla *Queue* las subcategorías que tiene la categoría procesada (en el ejemplo anterior son 2 las subcategorías a almacenar). En los registros almacenados en la tabla *Queue*, también se almacenará la información de quién es el padre de estas subcategorías para que, cuando haya que procesarlas, se almacene dicha información en la tabla *Hierarchies* (ver Ilustración 30, zona e).

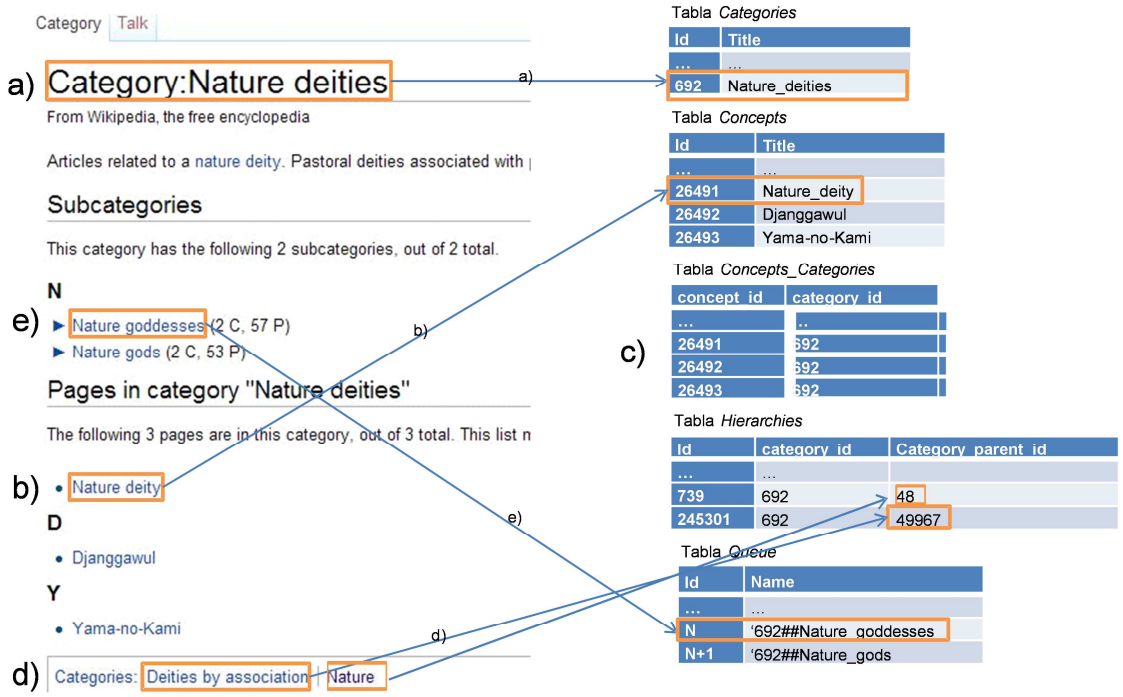

Ilustración 30: Ejemplo ilustrativo de almacenamiento de estructura de Wikipedia en la base de datos

El proceso terminará cuando no haya más categorías en la tabla *Queue* que procesar.

Por otro lado, tenemos la clase *SimilarityProcess*, que será la encargada del procesamiento de la estructura de Wikipedia almacenada en el proceso anterior para el cálculo de la similitud. Se encargará de gestionar los objetos *Pairs*, *Concepts* y *Categories*, del componente *Resources* (que veremos en el siguiente apartado) necesarios para el cálculo de las distintas medidas de similitud, a través de las clases *PathMeasures* (del componente *Measures*).

Los resultados obtenidos tras procesar las medidas se imprimen, bien por pantalla, bien en un archivo.

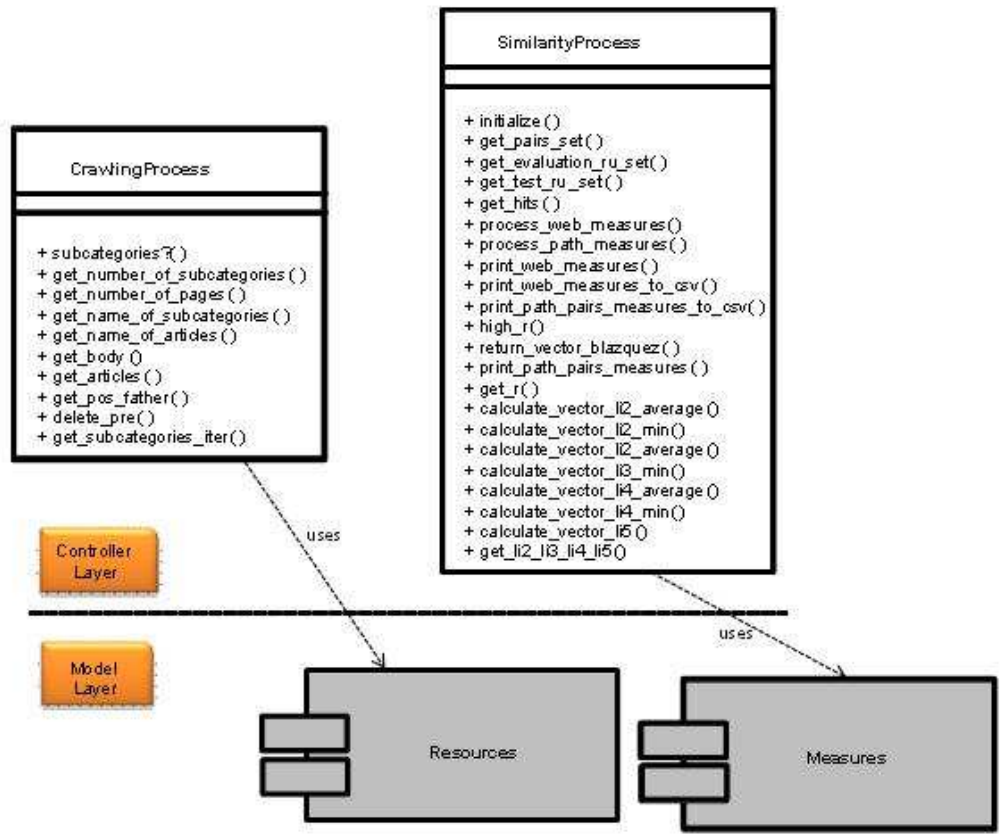

Ilustración 31: Diagrama de clases de la capa controladora

6.4.2. Capa de modelo

En esta sección veremos en más detalle los componentes de la capa de modelo (*Model layer*).

6.4.2.1. Componente *Measures*

Este componente está formado por las siguientes clases o módulos.

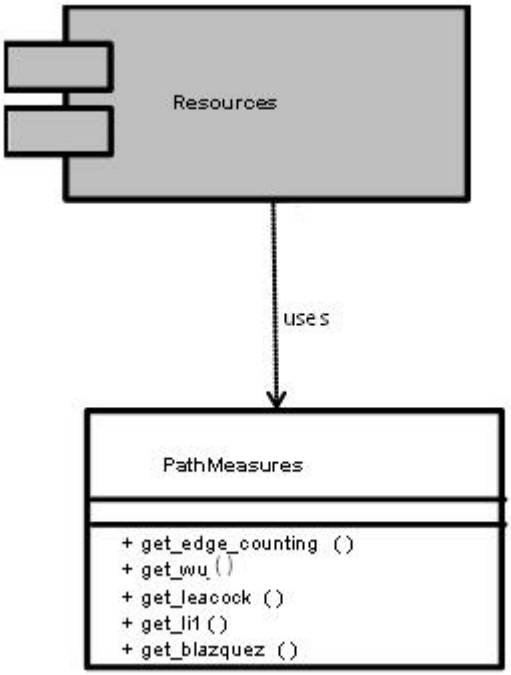

Ilustración 32: Diagrama de clases del componente Resources

Desde la clase *SimilarityProcess*, de la capa controladora, se lanza:

1) el cálculo de las medidas basadas en caminos.

Para cada medida se irá procesando uno a uno cada par del conjunto de pares a procesar; es decir, por cada instancia *Pair* que tengamos en memoria.

6.4.2.2. Componente *Resources*

En la Ilustración 33, podemos ver las diferentes clases que conforman o se relacionan con este componente.

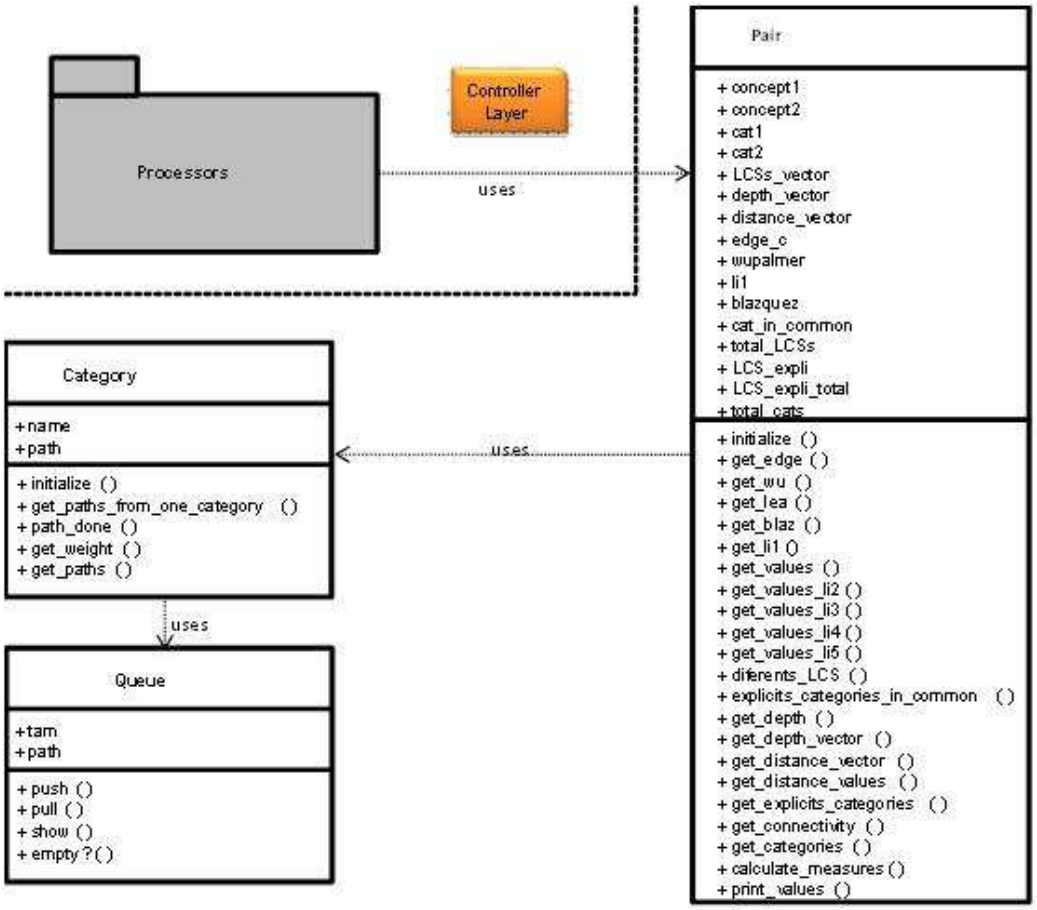

Ilustración 33: Componente *Resources* de la capa de modelo

Se almacenan en memoria los pares de conceptos a comparar, los valores correspondientes de los sets de (Rubenstein & Goodenough, 1965), y los *hits* (o número de resultados de una búsqueda) de las búsquedas basadas en Web, a partir de ficheros de texto. Para cada par de conceptos tendremos un objeto *Pair.* Dentro de este objeto, tendremos tantos objetos *Category* como categorías a las que pertenecen cada uno de los dos objetos de cada par. Para cada categoría, necesitamos de un objeto *Queue* para el procesado de los posibles caminos de cada categoría mientras estamos ejecutando el proceso de crawling (ver “6.4.1: Capa controladora”)

6.4.3. Capa de servicios técnicos

Finalmente, la capa de servicios técnicos está formada por los componentes *Persistence* y *External Access*, formadas a su vez por la clase *DataBase* y las clases implementadas *mysql, uri, open-uri* y *net/http* (ya implementada en el propio lenguaje Ruby).

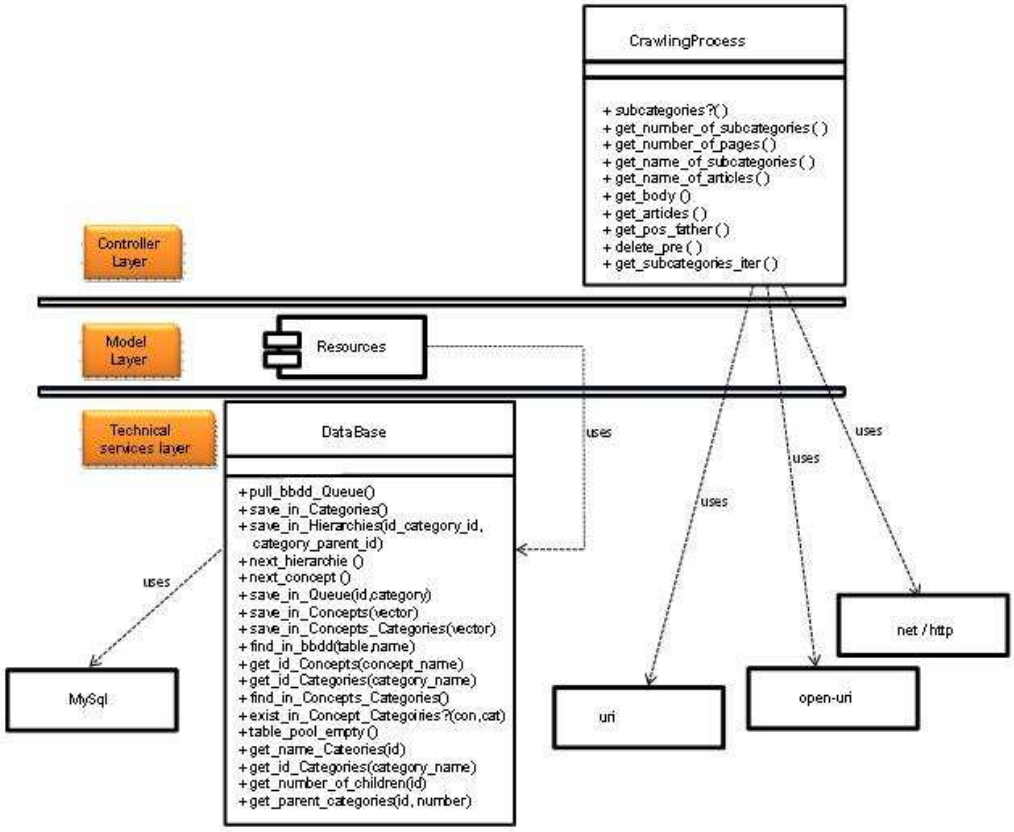

Ilustración 34: Capa de servicios técnicos

MySql proporcionará el almacenamiento en base de datos necesario, siendo gestionada dicha base de datos con la clase *DataBase*.

Las clases implementadas *uri, open-uri, net/http* ayudarán en el proceso de crawling a conectarse a las páginas web de Wikipedia necesarias para realizar su descarga.

7. ADAPTACIÓN DE MEDIDAS

En este capítulo explicaremos detalladamente el proceso de adaptación de las medidas tradicionales a Wikipedia, exponiendo los resultados obtenidos con cada una de ellas y comparándolas con las medidas tradicionales.

7.1. Adaptación de factores básicos

El objetivo del proyecto se centra en la adaptación de medidas tradicionales basadas en caminos, utilizando Wikipedia como taxonomía en lugar de WordNet o de cualquier otra. Para ello habrá que adaptarlas a la estructura de Wikipedia, resolviendo sus principales inconvenientes o dificultades de procesamiento que ya se comentaron en el apartado “4.5. Estructura de categorías de Wikipedia”, y que son principalmente la existencia de ciclos en el grafo y su estructura de herencia múltiple.

Para ello, adaptaremos cuatro factores importantes dentro de la estructura de Wikipedia, y que serán necesarios en el cálculo de diferentes fórmulas: 1) El cálculo del nodo común (*lcs)* entre dos conceptos; 2) el camino más corto entre dos conceptos; 3) la profundidad máxima de la jerarquía; y 4) la profundidad de un nodo. Para una mejor comprensión, tomaremos de referencia la Ilustración 35.

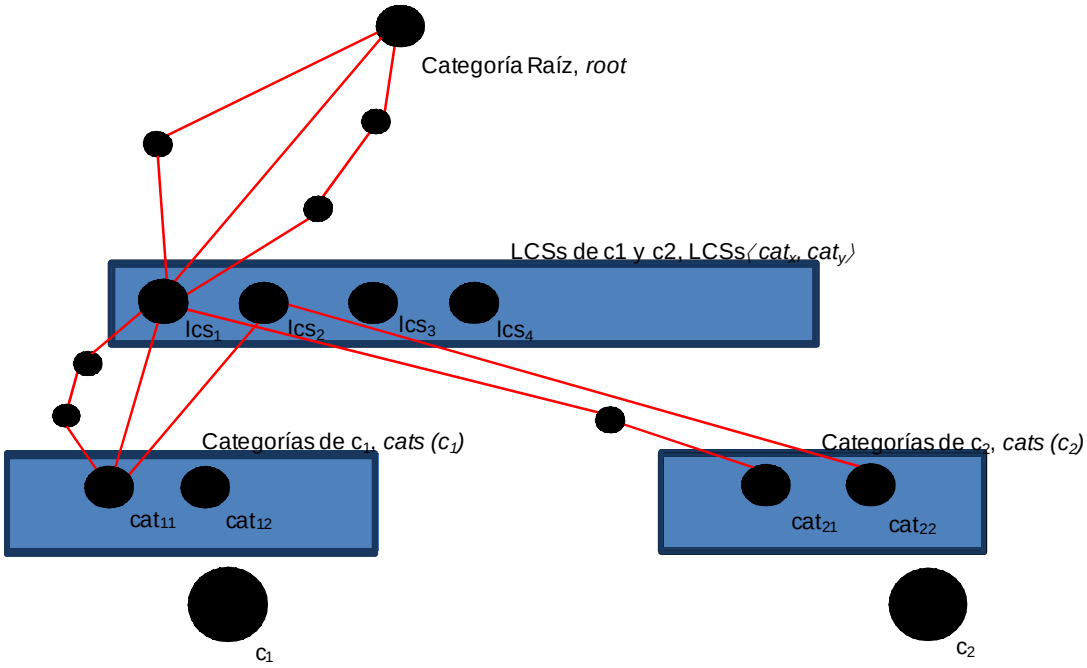

Ilustración 35: Ejemplo ilustrativo de profundidades y distancias entre en una taxonomía con herencia múltiple

En esta ilustración se puede ver que, si bien es cierto que la similitud se mide entre conceptos, en Wikipedia son asignados a una o más categorías, y será con esa estructura de categorías con la que trabajaremos.

7.1.1. Profundidad máxima de la jerarquía

El primer atributo a considerar es la *profundidad máxima* de la jerarquía, que llamaremos *d,* utilizada en algunas medidas tradicionales. En el caso de Wikipedia, en la versión descargada de enero del 2012, la profundidad máxima es de 20 arcos. Es decir, el camino máximo que hay que realizar para alcanzar la raíz desde una categoría cualquiera es de 20 saltos. Los bucles en caminos son eliminados de cualquier proceso de cómputo (el camino puede ser infinito si nos encontramos dentro de un bucle), como ya se comentó en el apartado "4.5. Estructura de categorías de Wikipedia".

7.1.2. Cálculo del nodo común entre dos conceptos

Dados dos conceptos c_1 y c_2, se extrae la lista de categorías a las que pertenecen, cats(c_1) y cats(c_2) respectivamente. Dadas estas listas, para cada par de categorías cruzadas $\langle cat_x, cat_y \rangle$, $cat_x \in cats(c_1)$, $cat_y \in cats(c_2)$, se calculan sus lcs's de manera tradicional, $LCSs(cat_x, cat_y)$. Se habla en plural debido a la estructura de herencia múltiple de Wikipedia. De esta manera, el nodo común entre los conceptos c_1 y c_2 es en realidad un vector de lcs's, $LCSs\langle c_1, c_2 \rangle$, tal que:

$$LCSs\,\langle c_1, c_2\rangle = \bigcup_{\forall\,\langle cat_x, cat_y\rangle} \{LCSs\langle cat_x, cat_y\rangle\}$$

Ecuación 25: Ecuación para el cálculo del vector LCSs entre dos conceptos

7.1.3. Camino más corto entre dos nodos

Uno de los atributos más usados en las medidas de similitud basadas en caminos es el *camino más* corto entre dos conceptos, a través del nodo común más cercano que los incluye (su *lcs)*. El camino más corto entre dos nodos en un grafo es el número de saltos (arcos o nodos intermedios) mínimos que los unen. Sin embargo, como ya hemos visto en el apartado anterior, debido a la estructura de herencia múltiple en la clasificación de conceptos de Wikipedia, tenemos que procesar más de un valor para la distancia, pues se cuenta con 1) diversos pares de categorías; y 2) múltiples caminos entre cada par.

Para ello, se introduce el vector *dist_vector*. Dicho vector tendrá tantos elementos como *lcs's* tenga el par de conceptos c_1 y c_2; es decir, los mismos

elementos que el vector LCSs⟨c_1,c_2⟩. El valor de cada dimensión del vector se calcula obteniendo la distancia mínima entre esos conceptos a través del lcs al que se refiere dicha dimensión, es decir:

$$dist\_vector\langle c1, c2\rangle: LCSs \rightarrow [0,2 \cdot d]$$

$$lcs \rightarrow \min_{\forall cat \in cats(c_1)} \{length(cat, lcs)\} + \min_{\forall cat \in cats(c_2)} \{length(cat, lcs)\}$$

Ecuación 26: Función de vector de distancia

Donde *length(cat,lcs)* es el camino más corto entre la categoría *cat* y la categoría *lcs*.

En el ejemplo de la Ilustración 35, vemos como el elemento lcs_1, se obtiene a partir de las categorías cat_{11} y cat_{21}. Hay varios caminos para llegar desde cat_{11} hasta cat_{21} pasando por su lcs_1, así que se selecciona el formado por menos saltos; es decir, el camino de cat_{11} a lcs_1 con un salto, y la distancia mínima entre lcs_1 y cat_{21} (2 saltos). Por tanto, *dist_vector⟨c1,c2⟩* tendrá un primer elemento con valor 3 (1+2).

LCSs ⟨c_1,c_2⟩	lcs_1	lcs_2	…	lcs_p
dist_vector ⟨c1,c2⟩	1 + 2 = 3		…	…

Ilustración 36: Ejemplo ilustrativo de dist_vector

Teniendo en cuenta la Ilustración 35, esta misma operación se realizará para el resto de *lcs's* conseguidos a través de las combinaciones de las categorías <cat_{11}, cat_{21}>, <cat_{11}, cat_{22}>, <cat_{12}, cat_{21}>, <cat_{12} cat_{22}>.

7.1.4. Profundidad de un nodo

La profundidad de un nodo suele utilizarse en las medidas tradicionales para calcular la profundidad de un *lcs*, obteniéndose contando el número de saltos entre ese nodo y la raíz. Debido a la propiedad de herencia múltiple de Wikipedia, además de tener múltiples *lcs's*, pueden existir varios valores de profundidad para cada *lcs* dado. En la Ilustración 35, puede verse como la profundidad de lcs_1 puede tener una distancia de 1, 2 ó 3 saltos.

Dada una categoría, una profundidad menor indicaría menos distancia a la raíz, pero también menor especialización. Por ello, inicialmente se tendrán en cuenta todos los posibles caminos a la raíz, no nos quedaremos con el que suponga la distancia mínima.

Para ello, y dados dos conceptos c_1 y c_2, contamos con otro vector, *depth_vector*$\langle c_1,c_2\rangle$, que también tendrá tantos elementos como *lcs's* tenga el par de conceptos c_1 y c_2. En este caso:

$$depth\_vector\langle c1, c2\rangle : LCSs\langle c1, c2\rangle \rightarrow ([0,2 \cdot d], [0,2 \cdot d], [0,2 \cdot d])$$

$$lcs \rightarrow (depth\_vector\langle c1, c2\rangle\ (lcs)_{min} = min\{paths(lcs, root)\},$$

$$depth\_vector\langle c1, c2\rangle\ (lcs)_{avg} = avg\{paths(lcs, root)\},$$

$$depth\_vector\langle c1, c2\rangle\ (lcs)_{max} = max\{paths(lcs, root)\})$$

Ecuación 27: Ecuación profundidad máxima

Donde *paths* (cat_1,cat_2) es el conjunto de longitudes de los posibles caminos que existen entre dos categorías cat_1 y cat_2.

De vuelta en la Ilustración 35, para el lcs_1 hay 3 posibles caminos hacia la raíz: uno con dos saltos, otro con un salto y otro con tres saltos (líneas de color rojo). Así pues, la primera dimensión de *depth_vector*$\langle c_1,c_2\rangle$*(*lcs_1*)* tendrá como valor una tripleta con los valores (1,2,3) (profundidad mínima, media y máxima):

LCSs $\langle c_1,c_2\rangle$	lcs_1	lcs_2	…	lcs_p
dist_vector ⟨c1,c2⟩	3	…	…	…
depth_vector ⟨c1,c2⟩	(1,2,3)	…	…	…

Ilustración 37: Ejemplo ilustrativo de *dist_vector* y *depth_vector*

7.2. Adaptación de medidas tradicionales a Wikipedia

Los siguientes apartados explican las distintas medidas tradicionales basadas en caminos que han sido adaptadas para Wikipedia.

7.2.1. Adaptación de la medida de (Rada et al., 1989)

Recordemos que la medida de (Rada et al., 1989) se centraba en el camino más corto entre dos conceptos (ver Ecuación 1). Esta medida se adapta a Wikipedia mediante la utilización de *dist_vector*. Para ello se obtiene el valor mínimo, máximo y medio de todos los valores de *dist_vector* entre un par de conceptos, obteniendo 3 variantes:

$$sim_{rada\_min}(c_1, c_2) = 2 \cdot d - min\{dist\_vector<c_1, c_2>\};$$

$$sim_{rada\_avg}(c_1, c_2) = 2 \cdot d - avg\{dist\_vector<c_1, c_2>\};$$

$$sim_{rada\_max}(c_1, c_2) = 2 \cdot d - max\{dist\_vector<c_1, c_2>\};$$

Ecuación 28: Medidas de similitud de (Rada et al., 1989) adaptadas

7.2.2. Adaptación de la medida de (Wu y Palmer, 1994)

La medida de (Wu & Palmer, 1994) utiliza como variables la distancia mínima entre los conceptos c_1 y c_2 a su *lcs* y la profundidad del *lcs* a la raíz (ver Ecuación 3), por lo que hacemos uso de los vectores *dist_vector* y *depth_vector.*

Como cada *lcs* tiene 3 posibles valores de profundidad (mínimo, medio y máximo), se calcula la medida por cada *lcs* para cada uno de esos 3 valores, obteniendo un nuevo vector llamado *wp_vector* (ver Ecuación 29).

$$wp\_vector<c_1, c_2> : LCSs(c1, c2) \rightarrow ([0,1], [0,1], [0,1])$$

$$lcs \rightarrow (wp\_vector(c1, c2)(lcs)_{min} = \frac{2 \cdot depth\_vector(c1, c2)(lcs)_{min}}{dist\_vector(c1, c2)(lcs) + 2 \cdot depth\_vector(c1, c2)(lcs)_{min}},$$

$$wp\_vector(c1, c2)\ (lcs)_{avg} = \frac{2 \cdot depth\_vector(c1, c2)(lcs)_{avg}}{dist\_vector(c1, c2)(lcs) + 2 \cdot depth\_vector(c1, c2)(lcs)_{avg}},$$

$$wp\_vector(c1, c2)\ (lcs)_{max} = \frac{2 \cdot depth_{vector(c1}, c2)(lcs)_{max}}{dist_{vector(c1}, c2)(lcs) + 2 \cdot depth_{vector(c1}, c2)(lcs)_{max}})$$

Ecuación 29: Medidas de similitud de (Wu y Palmer, 1994) adaptadas para cada lcs

LCSs(c_1,c_2)	lcs_1	lcs_2	…	lcs_p
dist_vector (c_1,c_2)	3	…	…	…
depth_vector (c_1,c_2)	(1,2,3)	…	…	…
wp_vector (c_1,c_2)	(2·1 / 3+2·1, 2·2 / 3+2·2, 2·3 / 3+2·3)	…	…	…

Ilustración 38: Vectores para medida adaptada de (Wu y Palmer, 1994)

Una vez obtenido este vector, se agrupan en 3 subconjuntos distintos los valores generados en el paso anterior con la profundidad media, mínima y máxima:

$$WP_{min}(c1,c2) = \bigcup_{lcs \in LCS(c1,c2)} \{wp\_vector(c1,c2)(lcs)_{min}\};$$

$$WP_{avg}(c1,c2) = \bigcup_{lcs \in LCS(c1,c2)} \{wp\_vector(c1,c2)(lcs)_{avg}\};$$

$$WP_{max}(c1,c2) = \bigcup_{lcs \in LCS(c1,c2)} \{wp\_vector(c1,c2)(lcs)_{max}\};$$

Ecuación 30: Subconjuntos de todos los valores de la medida (Wu y Palmer, 1994) adaptada a la profundidad mínima, media y máxima de los LCSs

Finalmente, con ayuda de esos 3 subconjuntos se originan 9 medidas de similitud distintas:

$$sim_{adapted\_wp\_\min\_min}\langle c_1, c_2\rangle = min\{WP_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\mathrm{avg}\_min}\langle c_1, c_2\rangle = min\{WP_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\max\_min}\langle c_1, c_2\rangle = min\{WP_{max}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\min\_avg}\langle c_1, c_2\rangle = avg\{WP_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\mathrm{avg}\_avg}\langle c_1, c_2\rangle = avg\{WP_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\max\_avg}\langle c_1, c_2\rangle = avg\{WP_{max}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\min\_max}\langle c_1, c_2\rangle = max\{WP_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\mathrm{avg}\_max}\langle c_1, c_2\rangle = max\{WP_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_wp\_\max\_max}\langle c_1, c_2\rangle = max\{WP_{max}\langle c_1, c_2\rangle\};$$

Ecuación 31: Medidas de similitud de (Wu y Palmer, 1994) adaptadas

7.2.3. Adaptación de la medida de (Leacock y Chodorow, 1994)

La medida de (Leacock & Chodorow, 1994) utilizaba la distancia mínima entre dos conceptos, *length(c_1,c_2),* y la profundidad máxima de la jerarquía, *d*. Se aplica la fórmula de la medida para cada valor de la distancia mínima de cada *lcs*, como se hizo con la medida de (Rada et al., 1989):

$$sim_{adapted\_lc\_min}(c_1, c_2) = \frac{-\log(min\{dist\_vector\langle c1, c2\rangle\})}{2 \cdot d};$$

$$sim_{adapted\_lc\_avg}(c_1, c_2) = \frac{-\log(avg\{dist\_vector\langle c1, c2\rangle\})}{2 \cdot d};$$

$$sim_{adapted\_lc\_\max}(c_1, c_2) = \frac{-\log(max\{dist\_vector\langle c1, c2\rangle\})}{2 \cdot d};$$

Ecuación 32: Medidas de similitud de (Leacock y Chodorow, 1994) adaptadas

7.2.4. Adaptación de la medida de (Blázquez-del-Toro et al., 2008)

Para adaptar la medida de (Blázquez-del-Toro et al., 2008), utilizamos *depth_vector*, una constante *k* (valores comprendidos entre 0,25 y 2,5, aumentando 0.25 cada vez), y el radio de información, que vendrá dado por los cocientes E_{lcs}/E_{c1} y E_{lcs}/E_{c2} (ver Ilustración 17). Para el cálculo de estos cocientes, necesitaremos en lugar de *dist_vector*, el vector *dist_min$_1$* y el vector *dist_min$_2$*. Estos vectores son el resultado de almacenar las distancias mínimas entre el c_1 y cada *lcs* en un vector, y las distancias mínimas entre cada *lcs* y el c_2 en otro vector. Si sumáramos los dos vectores obtendríamos el anteriormente explicado *dist_vector*. De esta manera, la medida original dará 3 medidas adaptadas por cada valor de la constante *k* (más tarde, en el "8.2.4. Resultados para la medida de (Blázquez-del-Toro et al., 2008)" se verá con qué valor de *k* se obtienen mejores resultados).

$$sim_{adapted\_bl\_\min}(c_1, c_2) = \frac{a_{min} * b_{min}}{a_{min} + b_{min} - (a_{min} * b_{min})};$$

$$sim_{adapted\_bl\_avg}(c_1, c_2) = \frac{a_{avg} * b_{avg}}{a_{avg} + b_{avg} - (a_{avg} * b_{avg})};$$

$$sim_{adapted\_bl\_\max}(c_1, c_2) = \frac{a_{max} * b_{max}}{a_{max} + b_{max} - (a_{max} * b_{max})};$$

$$a_{\mathrm{op}} = \frac{k * n_{op}}{k + n_{op} + \log\left(\frac{E_{lcs}}{E_{c1}}\right)}$$

$$b_{\mathrm{op}} = \frac{k * n_{op}}{k + n_{op} + \log\left(\frac{E_{lcs}}{E_{c2}}\right)}$$

Ecuación 33: Medidas de similitud de (Blazquez-del-Toro et al., 2008) adaptadas

Donde *op* es una de las tres operaciones *min, avg* y *max.*

7.2.5. Adaptación de la medida de (Li et al., 2003)

(Li et al., 2003) trabaja con varias fórmulas en función de varios factores. De sus cinco medidas basadas en caminos, cuatro de ellas utilizan constantes que serán necesarias calcular, por lo que su proceso de programación será distinto que la primera de sus medidas, que no necesitará de constantes.

Medida 1

$$sim_{adapted\_li1\_min}(c_1, c_2) = 2 \cdot d - min\{dist\_vector\langle c1, c2\rangle\};$$

$$sim_{adapted\_li1\_avg}(c_1, c_2) = 2 \cdot d - avg\{dist\_vector\langle c1, c2\rangle\};$$

$$sim_{adapted\_li1\_max}(c_1, c_2) = 2 \cdot d - max\{dist\_vector\langle c1, c2\rangle\};$$

Ecuación 34: Medidas de similitud para el método 1 de (Li et al., 2003) adaptadas

Esta medida es la misma que la que habíamos calculado para el método de (Rada et al., 1989). Es igual al doble de la profundidad máxima de la jerarquía, menos la distancia mínima de los conceptos. Debido a la existencia de herencia múltiple en Wikipedia, tendremos un valor de distancia mínima para cada *lcs* perteneciente a los dos conceptos en cuestión. Estos valores de distancia mínima vendrán dados por el vector *dist_vect*, ya calculado.

Medida 2

La segunda medida se basa en la combinación lineal del camino más corto entre los conceptos c_1 y c_2 y la profundidad del *lcs* entre c_1 y c_2.

Para ello, nuestra adaptación es una combinación lineal que utiliza la distancia mínima de cada *lcs* a través de *dist_vector*, ya usada en los métodos

anteriores, y el factor de la profundidad de cada *lcs* a través de *depth_vector*. Además, el método 2 de (Li et al., 2003) utiliza dos constantes que tendrán que ser calculadas.

Como cada *lcs* tiene 3 posibles valores de profundidad (mínimo, medio y máximo), se calcula la medida por cada *lcs* para cada uno de esos 3 valores, obteniendo un nuevo vector llamado *li2_vector* (ver Ecuación 35)

$$li2\_vector<c_1, c_2> : LCSs(c1, c2) \rightarrow ([0,1], [0,1], [0,1])$$

$$lcs \rightarrow (li2\_vector(c1, c2)(lcs)_{min} =$$

$$\alpha(2 \cdot d - \{dist\_vector(c1, c2)\}) + \beta(min\{depth\_vector\langle c_1, c_2\rangle(\text{lcs})_{min}\}),$$

$$li2\_vector(c1, c2)\ (lcs)_{avg} =$$

$$\alpha(2 \cdot d - \{dist\_vector(c1, c2)\}) + \beta\left(avg\left\{depth\_vector\langle c_1, c_2\rangle(\text{lcs})_{avg}\right\}\right),$$

$$li2\_vector(c1, c2)\ (lcs)_{max} =$$

$$\alpha(2 \cdot d - \{dist\_vector(c1, c2)\}) + \beta(max\{depth\_vector\langle c_1, c_2\rangle(\text{lcs})_{max}\})$$

Ecuación 35: Medidas de similitud de (Li et al., 2003) adaptadas para cada lcs (método 2)

LCSs(c_1,c_2)	lcs_1	lcs_2	…	lcs_p
dist_vector (c_1,c_2)	3	…	…	…
depth_vector (c_1,c_2)	(1,2,3)	…	…	…
li2_vector (c_1,c_2)	(α·1(2d-3)+β·1, α·1(2d-3)+β·2, α·1(2d-3)+β·3)	…	…	…

Ilustración 39: Vectores para medida de (Li et al., 2003), método 2

Una vez obtenido este vector, se agrupan en 3 subconjuntos distintos los valores generados en el paso anterior con la profundidad media, mínima y máxima:

$$LI2_{min}(c1, c2) = \bigcup_{lcs \in LCS(c1,c2)} \{li2\_vector(c1, c2)(lcs)_{min}\};$$

$$LI2_{avg}(c1,c2) = \bigcup_{lcs \in LCS(c1,c2)} \{li2\_vector(c1,c2)(lcs)_{avg}\};$$

$$LI2_{max}(c1,c2) = \bigcup_{lcs \in LCS(c1,c2)} \{li2\_vector(c1,c2)(lcs)_{max}\};$$

Ecuación 36: Subconjuntos de tosos los valores de la medida (Li et al., 2003; método 2) adaptada a la profundidad mínima, media y máxima de los LCSs

Finalmente, con ayuda de esos 3 subconjuntos se originan 9 medidas de similitud distintas:

$$sim_{adapted\_li2\_\min\_min}\langle c_1,c_2\rangle = min\{LI2_{min}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\mathrm{avg}\_min}\langle c_1,c_2\rangle = min\{LI2_{avg}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\max\_min}\langle c_1,c_2\rangle = min\{LI2_{max}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\min\_avg}\langle c_1,c_2\rangle = avg\{LI2_{min}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\mathrm{avg}\_avg}\langle c_1,c_2\rangle = avg\{LI2_{avg}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\max\_avg}\langle c_1,c_2\rangle = avg\{LI2_{max}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\min\_max}\langle c_1,c_2\rangle = max\{LI2_{min}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\mathrm{avg}\_max}\langle c_1,c_2\rangle = max\{LI2_{avg}\langle c_1,c_2\rangle\};$$

$$sim_{adapted\_li2\_\max\_max}\langle c_1,c_2\rangle = max\{LI2_{max}\langle c_1,c_2\rangle\};$$

Ecuación 37: Medidas de similitud adaptadas para el método 2 de (Li et al., 2003)

Medida 3

La tercera medida se basa en la combinación no lineal del camino más corto entre dos conceptos c_1 y c_2, así que de nuevo se adapta la fórmula a través de *dist_vector*:

$$sim_{adapted\_li3\_min}(c_1,c_2) = e^{-\alpha \cdot min\{dist\_vector(c1,c2)\}}$$

$$sim_{adapted\_li3\_avg}(c_1, c_2) = e^{-\alpha \cdot avg\{dist\_vector\langle c1,c2\rangle\}}$$

$$sim_{adapted\_li3\_max}(c_1, c_2) = e^{-\alpha \cdot max\{dist\_vector\langle c1,c2\rangle\}}$$

Ecuación 38: Medida de similitud adaptada para el método 3 de (Li et al., 2003)

Medida 4

La cuarta medida de (Li et al., 2003) se basa en la combinación no lineal de los factores distancia mínima entre dos conceptos c_1 y c_2 y profundidad del *lcs* entre c_1 y c_2.

Como cada *lcs* tiene 3 posibles valores de profundidad (mínimo, medio y máximo), se calcula la medida por cada *lcs* para cada uno de esos 3 valores, obteniendo un nuevo vector llamado *li4_vector* (ver Ecuación 39)

$$li4\_vector<c_1, c_2> : LCSs\langle c1, c2\rangle \rightarrow ([0,1],[0,1],[0,1])$$

$$lcs \rightarrow (li4\_vector\langle c1, c2\rangle (lcs)_{min} = e^{-\alpha \cdot \{dist\_vector\langle c1,c2\rangle\}} \cdot \frac{e^{\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}} - e^{-\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}}}{e^{\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}} + e^{-\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}}},$$

$$li4\_vector\langle c1, c2\rangle (lcs)_{avg} = e^{-\alpha \cdot \{dist\_vector\langle c1,c2\rangle\}} \cdot \frac{e^{\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}} - e^{-\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}}}{e^{\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}} + e^{-\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}}},$$

$$li4\_vector\langle c1, c2\rangle (lcs)_{max} = e^{-\alpha \cdot \{dist\_vector\langle c1,c2\rangle\}} \cdot \frac{e^{\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}} - e^{-\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}}}{e^{\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}} + e^{-\beta \cdot \{depth\_vector\ \langle c_1,c_2\rangle\}}})$$

Ecuación 39: Medidas de similitud de (Li et al., 2003; método 4) adaptadas para cada *lcs*

LCSs(c_1,c_2)	lcs_1	lcs_2	…	lcs_p
dist_vector $\langle c_1,c_2\rangle$	3	…	…	…
depth_vector $\langle c_1,c_2\rangle$	(1,2,3)	…	…	…
li4_vector $\langle c_1,c_2\rangle$	$= (e^{-\alpha \cdot \{3\}} \cdot \frac{e^{\beta \cdot 1} - e^{-\beta \cdot 1}}{e^{\beta \cdot 1} + e^{-\beta \cdot 1}},$ $e^{-\alpha \cdot \{3\}} \cdot \frac{e^{\beta \cdot 2} - e^{-\beta \cdot 2}}{e^{\beta \cdot 2} + e^{-\beta \cdot 2}},$	…	…	…

$e^{-\alpha \cdot \{3\}} \cdot \frac{e^{\beta \cdot 3} - e^{-\beta \cdot 3}}{e^{\beta \cdot 3} + e^{-\beta \cdot 3}})$			

Ilustración 40: Vectores para medida de (Li et al., 2003), método 4

Una vez obtenido este vector, se agrupan en 3 subconjuntos distintos los valores generados en el paso anterior con la profundidad media, mínima y máxima:

$$LI4_{min}\langle c1, c2\rangle = \bigcup_{lcs \in LCS\langle c1,c2\rangle} \{li4\_vector\langle c1, c2\rangle(lcs)_{min}\};$$

$$LI4_{avg}\langle c1, c2\rangle = \bigcup_{lcs \in LCS\langle c1,c2\rangle} \{li4\_vector\langle c1, c2\rangle(lcs)_{avg}\};$$

$$LI4_{max}\langle c1, c2\rangle = \bigcup_{lcs \in LCS\langle c1,c2\rangle} \{li4\_vector\langle c1, c2\rangle(lcs)_{max}\};$$

Ecuación 40: Subconjuntos de tosos los valores de la medida (Li et al., 2003; método 4) adaptada a la profundidad mínima, media y máxima de los LCSs

Finalmente, y con la ayuda de esos 3 subconjuntos se originan 9 medidas de similitud distintas:

$$sim_{adapted\_li4\_\min\_min}\langle c_1, c_2\rangle = min\{LI4_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\mathrm{avg}\_min}\langle c_1, c_2\rangle = min\{LI4_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\max\_min}\langle c_1, c_2\rangle = min\{LI4_{max}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\min\_avg}\langle c_1, c_2\rangle = avg\{LI4_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\mathrm{avg}\_avg}\langle c_1, c_2\rangle = avg\{LI4_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\max\_avg}\langle c_1, c_2\rangle = avg\{LI4_{max}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\min\_max}\langle c_1, c_2\rangle = max\{LI4_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\mathrm{avg}\_max}\langle c_1, c_2\rangle = max\{LI4_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li4\_\max\_max}\langle c_1, c_2\rangle = max\{LI4_{max}\langle c_1, c_2\rangle\};$$

Ecuación 41: Medidas de similitud para el método 4 de (Li et al., 2003) adaptadas

Medida 5

Esta medida se basaba en la función no lineal de la profundidad de un *lcs* entre dos conceptos c_1 y c_2.

Como cada *lcs* tiene 3 posibles valores de profundidad (mínimo, medio y máximo), se calcula la medida por cada *lcs* para cada uno de esos 3 valores, obteniendo un nuevo vector llamado *li5_vector*.

$$li5\_vector<c_1, c_2>: LCSs\langle c1, c2\rangle \rightarrow ([0,1],[0,1],[0,1])$$

$$lcs \rightarrow (li5_{vector\langle c1}, c2\rangle (lcs)_{min} = \frac{e^{\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{min}\}} - e^{-\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{min}\}}}{e^{\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{min}\}} + e^{-\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{min}\}}},$$

$$li5\_vector\langle c1, c2\rangle\ (lcs)_{avg} = \frac{e^{\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{avg}\}} - e^{-\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{avg}\}}}{e^{\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{avg}\}} + e^{-\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{avg}\}}},$$

$$li5\_vector\langle c1, c2\rangle\ (lcs)_{max} = \frac{e^{\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{max}\}} - e^{-\beta\cdot\{depth\_vector\ \langle c_1,c_2\max\}}}{e^{\beta\cdot\{depth\_vector\ \langle c_1,c_2\rangle (lcs)_{max}\}} + e^{-\beta\cdot\{depth\_vector\ \langle c_1,c_2\max\}}})$$

Ecuación 42: Medidas de similitud de (Li et al., 2003) adaptadas para cada lcs (método 5)

LCSs(c_1,c_2)	lcs_1	lcs_2	…	lcs_p
depth_vector $\langle c_1,c_2\rangle$	(1,2,3)	…	…	…
li5_vector $\langle c_1,c_2\rangle$	$(\frac{e^{\beta\cdot 1} - e^{-\beta\cdot 1}}{e^{\beta\cdot 1} + e^{-\beta\cdot 1}}, \frac{e^{\beta\cdot 2} - e^{-\beta\cdot 2}}{e^{\beta\cdot 2} + e^{-\beta\cdot 2}}, \frac{e^{\beta\cdot 3} - e^{-\beta\cdot 3}}{e^{\beta\cdot 3} + e^{-\beta\cdot 3}})$	…	…	…

Ilustración 41: Vectores para medida de (Li et al., 2003), método 5

Una vez obtenido este vector, se agrupan en 3 subconjuntos distintos los valores generados en el paso anterior con la profundidad media, mínima y máxima.

$$LI5_{min}\langle c1, c2\rangle = \bigcup_{lcs\ \in LCS\langle c1,c2\rangle} \{li5\_vector\langle c1, c2\rangle (lcs)_{min}\};$$

$$LI5_{avg}\langle c1, c2\rangle = \bigcup_{lcs\ \in LCS\langle c1,c2\rangle} \{li5\_vector\langle c1, c2\rangle (lcs)_{avg}\};$$

$$LI5_{max}(c1, c2) = \bigcup_{lcs \in LCS(c1,c2)} \{li5\_vector(c1, c2)(lcs)_{max}\};$$

Ecuación 43: Subconjuntos de todos los valores de la medida (Li et al., 2003; método 5) adaptada a la profundidad mínima, media y máxima de los LCSs

Finalmente, con ayuda de esos 3 subconjuntos se originan 9 medidas de similitud distintas:

$$sim_{adapted\_li5\_min\_min}\langle c_1, c_2\rangle = min\{LI5_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_avg\_min}\langle c_1, c_2\rangle = min\{LI5_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_max\_min}\langle c_1, c_2\rangle = min\{LI5_{max}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_min\_avg}\langle c_1, c_2\rangle = avg\{LI5_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_avg\_avg}\langle c_1, c_2\rangle = avg\{LI5_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_max\_avg}\langle c_1, c_2\rangle = avg\{LI5_{max}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_min\_max}\langle c_1, c_2\rangle = max\{LI5_{min}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_avg\_max}\langle c_1, c_2\rangle = max\{LI5_{avg}\langle c_1, c_2\rangle\};$$

$$sim_{adapted\_li5\_max\_max}\langle c_1, c_2\rangle = max\{LI5_{max}\langle c_1, c_2\rangle\};$$

Ecuación 44: Medida de similitud adaptada para el método 5 de (Li et al., 2003)

Tendremos pues tres resultados en función del valor de profundidad (mínimo, medio y máximo).

8. EXPERIMENTOS Y RESULTADOS

En este apartado mostramos las pruebas realizadas para la obtención de nuestras medidas adaptadas junto con los resultados obtenidos. Dichos resultados serán comparados con los que en su día obtuvieron los autores de las medidas clásicas. Recordemos que los conceptos usados para el entrenamiento y la prueba se encuentran descritos en el apartado 6.3.

En este capítulo, a la hora de realizar comparaciones, utilizaremos los valores originales publicados de los artículos que se mencionan, para el conjunto de prueba.

Sin embargo, como no hay constancia de los resultados obtenidos para esos trabajos para el conjunto de entrenamiento, hemos calculado el valor de cada medida mediante una herramienta llamada Semantic Similarity System[5].

8.1. Secuencia de ejecución

A continuación se muestra el algoritmo utilizado para la obtención de nuestras medidas, junto una breve explicación. También se explica en más detalle el proceso seguido para la obtención de cada medida particular.

- *Crear objeto 'SimilarityProcess'*
  Este objeto contendrá varios vectores donde se recogerán los pares de conceptos, los valores numéricos de esos pares, los valores de similitud tras ser calculados y los índices de correlación. También contiene los métodos para cargar en memoria los valores y llamar a los procedimientos que calcularán las medidas basadas en caminos.
- *Cargar conjunto de pares de conceptos*
  Procedimiento de SimilarityProcess para cargar en memoria los conceptos de pares de conjunto de entrenamiento y prueba, leyéndolos desde archivo.
- *Cargar los valores numéricos de los pares de conceptos*
  Procedimiento de SimilarityProcess para cargar en memoria los valores de los pares de palabras de los conjuntos otorgados por humanos en (Rubenstein & Goodenough, 1965), leyéndolos desde archivo.
- *Procesar las medidas basadas en caminos*
  - Para cada par:
    - Crear objeto *Pair*
    - Obtener categorías para el par de conceptos

[5] Semantic Similarity System, de la Universidad Técnica de Creta, http://www.intelligence.tuc.gr/similarity/index.php

- Hallar todas las categorías de cada uno de los conceptos.
- Hallar los lcs's para cada par de conceptos cruzados.
- Obtener el vector de profundidad, *depth_vector*
- Obtener el vector de distancia, *distance_vector*

▪ Calcular medidas tradicionales a partir de los vectores (excepto li2, li3, li4 y li5, ya que estas medidas se calculan de modo diferente al depender de constantes que habrá que calcular. Se calculará el valor de la medida para cada posible valor de las constantes, y nos quedaremos con los valores que maximicen los coeficientes de correlación)

▪ Retornar valores

o Obtener medidas para li2, li3, li4, li5.

o Calcular coeficientes de correlación r (entre las medidas y los coeficientes) para todas las medidas

El coeficiente de correlación que se va a utilizar es el coeficiente de correlación de Pearson, cuya fórmula es la siguiente:

$$r_{xy} = \frac{n \cdot \sum x_i \cdot y_i - \sum x_i \cdot y_i}{\sqrt{n \cdot \sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n \cdot \sum y_i^2 - (\sum y_i)^2}}$$

Ecuación 45: Fórmula de coeficiente de correlación de Pearson sobre un muestra estadística

Dónde *n* sería el número de elementos de cada muestra, x_i los valores obtenidos por el experimento de (Rubenstein & Goodenough, 1965), e y_i los valores que obtenemos de cada una de las medidas.

8.2. Resultados

En este apartado mostramos los resultados obtenidos para cada medida adaptada junto con la comparación de la correspondiente medida tradicional.

8.2.1. Resultados de valores para las medidas adaptadas de (Rada et al., 1989)

Comenzamos mostrando los valores del conjunto de entrenamiento, teniendo en cuenta que obtenemos tres columnas en función de si el valor se obtiene usando la medida sim_{rada_min}, sim_{rada_avg} ó sim_{rada_max}.

Tabla 11: Valores de similitud con las medidas de (Rada & et al, 1989) adaptadas, conjunto de entrenamiento

Concepto1-Concepto2	Valor humano	rada_min	rada_avg	rada_max
Psychiatric_hospital-Cemetery (1)	0,79	34	32	29
Psychiatric_hospital-Fruit (2)	0,19	34	32	29
Psychiatric_hospital-Monk (3)	0,39	34	32	30
Autograph-Shore (4)	0,06	34	33	31
Autograph-Signature (5)	3,59	36	34	31
Automobile-Magician_(fantasy) (6)	0,11	29	28	26
Automobile-Cushion (7)	0,97	32	30	28
Bird-Woodland (8)	1,24	35	33	31
Boy-Rooster (9)	0,44	34	33	31
Boy-Philosophy (10)	0,96	35	35	33
Cemetery-Mound (11)	1,69	37	33	29
Cemetery-Graveyard (12)	3,88	40	36	30
Cemetery-Woodland (13)	1,18	38	34	30
Rope-Rope (14)	3,41	40	35	30
Rooster-Rooster (15)	3,68	40	34	28
Crane_(bird)-Rooster (16)	1,41	32	30	28
Cushion-Jewellery (17)	0,45	35	34	32
Cushion-Pillow (18)	3,84	40	36	32
Forest-Woodland (19)	3,65	40	35	30
Fruit-Furnace (20)	0,05	35	34	32
Furnace-Tool (21)	1,37	38	34	29
Glass-Jewellery (22)	1,78	39	35	30
Glass-Glass (23)	3,45	40	35	30
Graveyard-Psychiatric_hospital (24)	0,42	34	32	29
Smile-Tool (25)	0,18	33	32	30
Smile-Boy (26)	0,88	34	34	32
Smile-Smile (27)	3,46	40	36	32
Hill-Mound (28)	3,29	32	31	29
Hill-Woodland (29)	1,48	33	32	30
Magician_(fantasy)-Oracle (30)	1,82	34	32	28
Mound-Stove (31)	0,14	34	32	30
Mound-Shore (32)	0,97	33	32	30
Oracle-Philosophy (33)	2,61	38	34	31
Philosophy-Magician_(fantasy) (34)	2,46	35	33	30
Serfdom-Slavery (35)	3,46	38	35	32
Shore-Travel (36)	1,22	34	33	31
Shore-Woodland (37)	0,9	34	33	31

La siguiente tabla se obtiene usando las mismas medidas pero para el conjunto de prueba.

Tabla 12: Valores de similitud obtenidos con las medidas de (Rada et al, 1989) adaptadas, conjunto de prueba

Concepto1-Concepto2	Valor humano	rada_min	rada_avg	rada_max

Psychiatric_hospital-Psychiatric_hospital (1)	3,04	40	34	28
Automobile-Automobile (2)	3,92	40	32	24
Bird-Rooster (3)	2,63	33	32	31
Bird-Crane_(bird) (4)	2,63	37	31	25
Boy-Boy (5)	3,82	40	36	32
Sibling-Boy (6)	2,41	39	36	33
Brother_(Catholic)-Monk (7)	2,74	40	35	31
Automobile-Travel (8)	1,55	32	30	28
Rope-Smile (9)	0,02	34	33	31
Coast-Forest (10)	0,85	33	32	31
Coast-Hill (11)	1,26	38	34	30
Coast-Shore (12)	3,6	40	35	31
Crane_(machine)-Tool (13)	2,37	38	33	27
Food-Fruit (14)	2,69	39	35	32
Food-Rooster (15)	1,09	37	34	31
Forest-Graveyard (16)	1	38	34	31
Furnace-Stove (17)	3,11	37	33	28
Jewellery-Jewellery (18)	3,94	40	36	32
Glass-Magician_(fantasy) (19)	0,44	31	31	29
Tool-Tool (20)	3,66	40	35	28
Travel-Travel (21)	3,58	40	37	32
Boy-Magician_(fantasy) (22)	0,99	33	32	30
Magician_(fantasy)-Magician_(fantasy) (23)	3,21	40	34	28
Noon-Noon (24)	3,94	40	34	28
Monk-Oracle (25)	0,91	38	34	30
Monk-Slavery (26)	0,57	36	34	32
Noon-Rope (27)	0,04	31	31	30
Rooster-Travel (28)	0,04	33	32	31

Podemos representar dentro de una gráfica los valores, para ver su relación de modo visual y calcular así su coeficiente de correlación con los valores humanos.

Para que tengan la misma escala y se puedan comparar los valores humanos, que van entre 0 y 4, y los valores de nuestras medidas que van entre 0 y 40, multiplicamos la serie de valores humanos por la constante 40/4. El resultado del gráfico es el siguiente:

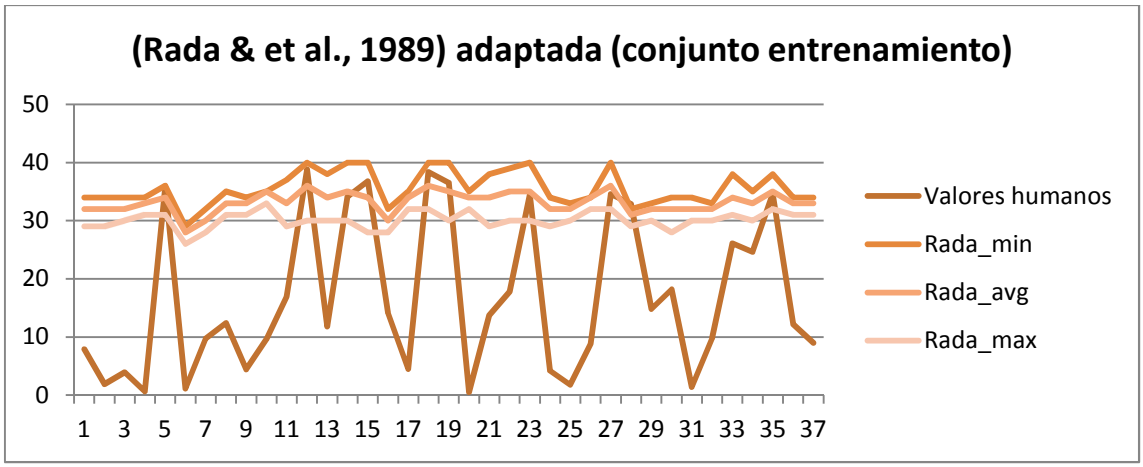

Ilustración 42: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Rada et al., 1989), conjunto de entrenamiento.

Calculamos el índice de correlación de cada conjunto con los valores humanos, y los comparamos con la correlación de la medida original.

Tabla 13: Coeficientes de correlación para las medidas de (Rada & et al, 1989) adaptada, conjunto de entrenamiento.

(Rada et al., 1989)	Rada_min	Rada_avg	Rada_max
0,550	0,721	0,572	0,094

A continuación realizamos el mismo proceso para el conjunto de prueba.

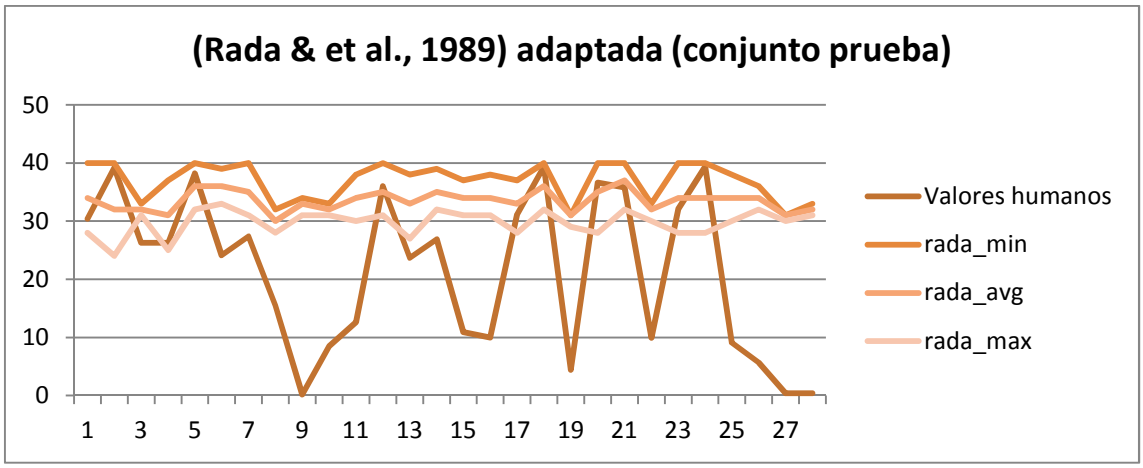

Ilustración 43: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Rada & et al, 1989), conjunto de prueba.

Ahora se calcula el índice de correlación para el conjunto de prueba.

Tabla 14: Coeficientes de correlación para las medidas de (Rada et al., 1989) adaptadas, conjunto de prueba

(Rada et al., 1989)	Rada_min	Rada_avg	Rada_max
0,6645	0,785	0,544	0,245

Como se puede ver, nuestra medida sim_{rada_min}, es la que ofrece los mejores resultados, aumentando de manera apreciable a la publicada originalmente en el conjunto de entrenamiento y aumentando más de una décima el índice de correlación (de 0,66 a 0,78) en el conjunto de prueba.

8.2.2. Resultados para las medidas adaptadas de (Wu & Palmer, 1994)

Presentamos a continuación los valores para las medidas adaptadas de (Wu & Palmer, 1994). Por motivos de espacio y presentación, los pares vienen representados por números del 1 al 37, en caso del conjunto de entrenamiento, y del 1 al 28, en caso del conjunto de prueba. Las columnas de letras de la A a la I representan las adaptaciones de la medida (A: $sim_{wp\_min\_min}$ B: $sim_{wp\_avg\_min}$, C: $sim_{wp\_min\_max}$, D: $sim_{wp\_avg\_min}$, E: $sim_{wp\_avg\_avg}$, F: $sim_{wp\_avg\_max}$, G: $sim_{wp\_max\_min}$, H: $sim_{wp\_max\_avg}$, I: $sim_{wp\_max\_max}$). La columna con las siglas 'Rb' representa los valores humanos.

Tabla 15: Valores obtenidos para las medidas de (Wu & Palmer, 1994) adaptadas, conjunto de entrenamiento

Par	Rb	A	B	C	D	E	F	G	H	I
1	0,79	0,266	0,343	0,4	0,421	0,519	0,625	0,56	0,644	0,75
2	0,19	0,266	0,378	0,5	0,444	0,531	0,625	0,6	0,641	0,72
3	0,39	0,285	0,346	0,428	0,444	0,523	0,571	0,583	0,65	0,692
4	0,06	0,307	0,367	0,5	0,5	0,546	0,6	0,631	0,67	0,727
5	3,59	0,307	0,438	0,6	0,47	0,598	0,736	0,608	0,699	0,8
6	0,11	0,222	0,247	0,266	0,38	0,435	0,521	0,518	0,555	0,621
7	0,97	0,25	0,29	0,333	0,421	0,479	0,521	0,56	0,596	0,621
8	1,24	0,307	0,392	0,545	0,47	0,555	0,666	0,608	0,666	0,737
9	0,44	0,307	0,351	0,4	0,47	0,557	0,6	0,608	0,671	0,72
10	0,96	0,363	0,457	0,545	0,533	0,626	0,705	0,666	0,733	0,783
11	1,69	0,285	0,443	0,727	0,444	0,59	0,769	0,583	0,687	0,8
12	3,88	0,285	0,629	1	0,444	0,729	1	0,583	0,796	1
13	1,18	0,285	0,463	0,8	0,444	0,609	0,833	0,583	0,707	0,857
14	3,41	0,285	0,634	1	0,444	0,703	1	0,583	0,764	1
15	3,68	0,285	0,553	1	0,444	0,643	1	0,5	0,717	1
16	1,41	0,21	0,241	0,315	0,347	0,408	0,461	0,461	0,527	0,667
17	0,45	0,333	0,454	0,714	0,5	0,593	0,714	0,636	0,692	0,727
18	3,84	0,333	0,61	1	0,5	0,715	1	0,636	0,789	1
19	3,65	0,307	0,597	1	0,47	0,687	1	0,583	0,754	1
20	0,05	0,333	0,46	0,615	0,533	0,604	0,666	0,666	0,721	0,783
21	1,37	0,307	0,497	0,8	0,47	0,621	0,857	0,56	0,71	0,889
22	1,78	0,307	0,348	0,428	0,5	0,546	0,6	0,6	0,658	0,933
23	3,45	0,285	0,618	1	0,444	0,707	1	0,545	0,762	1
24	0,42	0,266	0,343	0,4	0,421	0,519	0,625	0,56	0,644	0,75

25	0,18	0,285	0,367	0,461	0,5	0,537	0,6	0,631	0,655	0,696
26	0,88	0,333	0,393	0,5	0,5	0,573	0,625	0,636	0,694	0,75
27	3,46	0,333	0,624	1	0,5	0,7	1	0,6	0,765	1
28	3,29	0,266	0,362	0,5	0,444	0,515	0,555	0,583	0,624	0,667
29	1,48	0,285	0,35	0,461	0,444	0,517	0,588	0,583	0,637	0,667
30	1,82	0,25	0,376	0,571	0,4	0,522	0,631	0,538	0,628	0,72
31	0,14	0,285	0,367	0,428	0,444	0,523	0,571	0,583	0,646	0,667
32	0,97	0,285	0,391	0,533	0,47	0,542	0,588	0,608	0,646	0,696
33	2,61	0,307	0,47	0,75	0,47	0,62	0,833	0,608	0,721	0,889
34	2,46	0,285	0,394	0,5	0,47	0,561	0,705	0,6	0,67	0,783
35	3,46	0,333	0,493	0,833	0,5	0,639	0,875	0,636	0,733	0,9
36	1,22	0,307	0,367	0,5	0,5	0,546	0,6	0,631	0,67	0,727
37	0,9	0,307	0,378	0,5	0,47	0,548	0,625	0,608	0,665	0,7

Presentamos ahora el conjunto de prueba:

Tabla 16: Valores obtenidos para las medidas de (Wu & Palmer, 1994) adaptadas, conjunto de prueba

Par	Rb	A	B	C	D	E	F	G	H	I
1	3,04	0,25	0,569	1	0,4	0,651	1	0,538	0,72	1
2	3,92	0,2	0,533	1	0,363	0,599	1	0,5	0,665	1
3	2,63	0,307	0,348	0,461	0,47	0,537	0,588	0,6	0,652	0,696
4	2,63	0,21	0,458	0,823	0,347	0,55	0,842	0,444	0,627	0,857
5	3,82	0,333	0,611	1	0,5	0,706	1	0,636	0,775	1
6	2,41	0,363	0,545	0,857	0,533	0,682	0,888	0,666	0,779	0,909
7	2,74	0,307	0,539	1	0,5	0,655	1	0,636	0,744	1
8	1,55	0,25	0,303	0,428	0,421	0,484	0,555	0,56	0,597	0,636
9	0,02	0,307	0,378	0,5	0,47	0,548	0,625	0,608	0,665	0,692
10	0,85	0,307	0,366	0,461	0,47	0,535	0,6	0,6	0,654	0,692
11	1,26	0,285	0,52	0,857	0,444	0,614	0,857	0,545	0,692	0,875
12	3,6	0,307	0,592	1	0,47	0,676	1	0,571	0,746	1
13	2,37	0,235	0,484	0,857	0,434	0,592	0,857	0,56	0,68	0,889
14	2,69	0,333	0,545	0,888	0,5	0,666	0,909	0,666	0,755	0,933
15	1,09	0,307	0,477	0,727	0,47	0,618	0,769	0,608	0,707	0,824
16	1	0,307	0,47	0,8	0,47	0,616	0,833	0,608	0,714	0,857
17	3,11	0,333	0,429	0,666	0,454	0,586	0,727	0,538	0,688	0,769
18	3,94	0,333	0,647	1	0,5	0,723	1	0,636	0,791	1
19	0,44	0,266	0,291	0,307	0,444	0,49	0,571	0,545	0,61	0,666
20	3,66	0,25	0,575	1	0,454	0,673	1	0,583	0,747	1
21	3,58	0,333	0,685	1	0,5	0,761	1	0,636	0,818	1
22	0,99	0,285	0,348	0,428	0,444	0,526	0,6	0,583	0,653	0,692
23	3,21	0,25	0,545	1	0,4	0,635	1	0,538	0,705	1
24	3,94	0,25	0,574	1	0,4	0,661	1	0,5	0,731	1
25	0,91	0,285	0,51	0,833	0,444	0,629	0,833	0,583	0,717	0,85
26	0,57	0,333	0,432	0,5	0,5	0,596	0,666	0,636	0,707	0,75
27	0,04	0,285	0,316	0,375	0,444	0,491	0,571	0,571	0,62	0,666
28	0,04	0,307	0,374	0,533	0,47	0,554	0,608	0,608	0,66	0,696

Representamos dentro de dos gráficas los valores anteriores, para ambos conjuntos. Cada gráfica se separa en 3 submedidas por motivos de legibilidad.

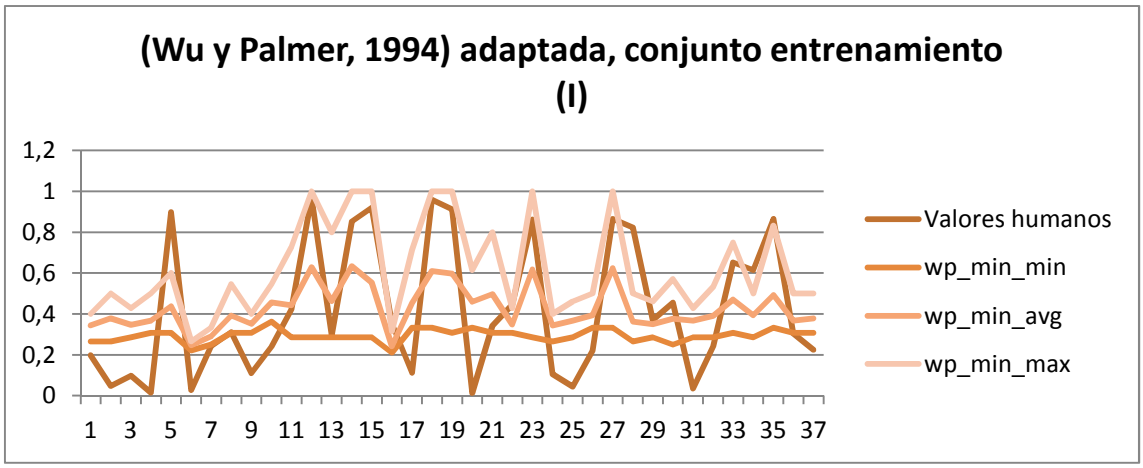

Ilustración 44: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Wu & Palmer, 1994), que utilizan los valores de profundidad mínimos de cada LCS, conjunto de entrenamiento

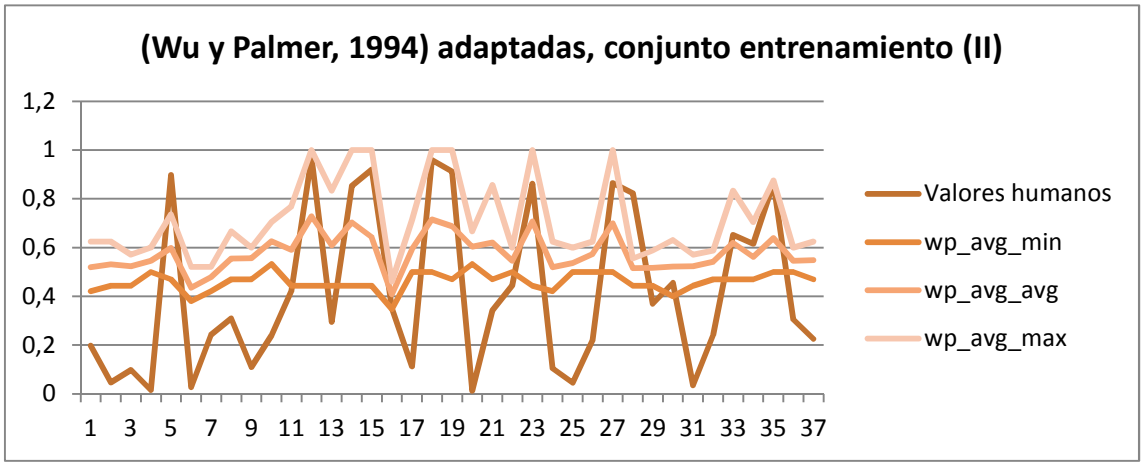

Ilustración 45: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Wu & Palmer, 1994), que utilizan los valores de profundidad medios de cada lcs, conjunto de entrenamiento

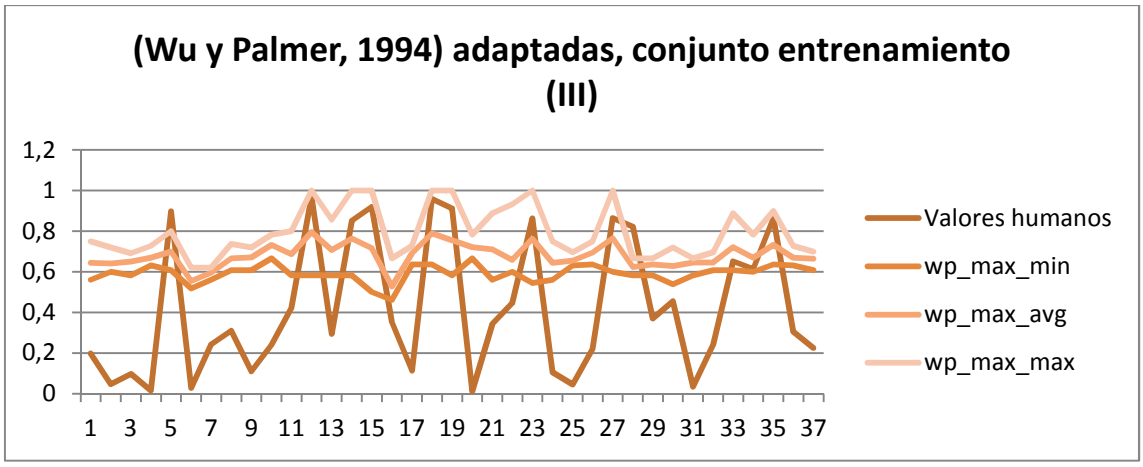

Ilustración 46: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Wu & Palmer, 1994), que utilizan los valores de profundidad máximos de cada lcs, conjunto de entrenamiento

Calculamos el índice de correlación, y comparamos con la correlación de la medida original.

Tabla 17: Coeficientes de correlación para las medidas de (Wu & Palmer, 1994) adaptadas, conjunto de entrenamiento

(Wu y Palmer, 1994)	Profundidades	sim_{wp_min}	sim_{wp_avg}	sim_{wp_max}
0,810	Mínimas	0,154	0,750	0,783
0,810	Medias	0,031	0,711	0,787
0,810	Máximas	0,199	0,631	0,763

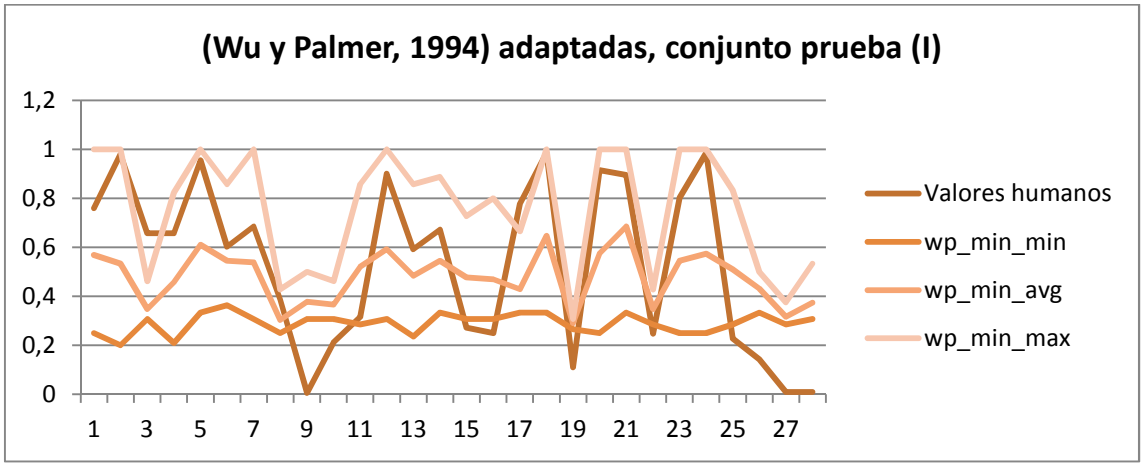

Ilustración 47: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Wu & Palmer, 1994), que utilizan los valores de profundidad mínimos de cada lcs, conjunto de prueba

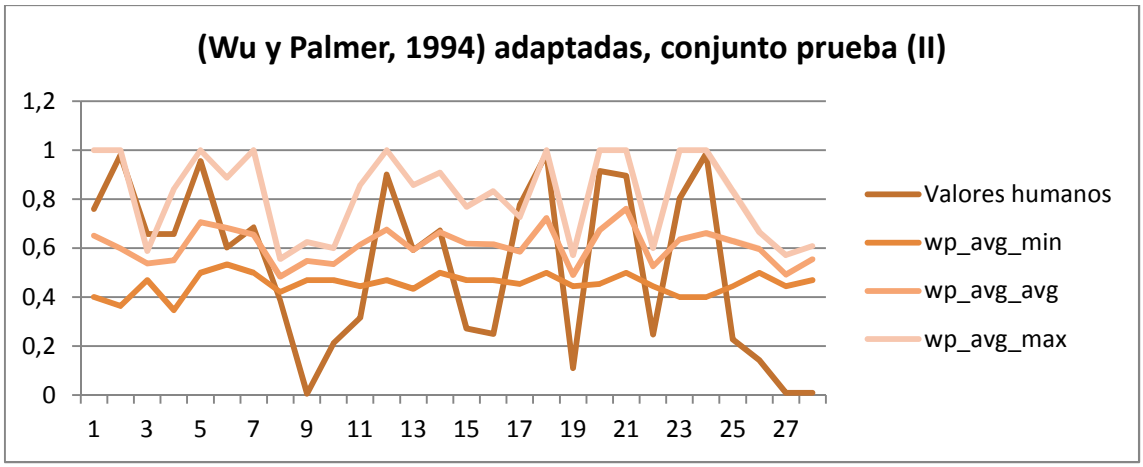

Ilustración 48: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Wu & Palmer, 1994), que utilizan los valores de profundidad medios de cada lcs, conjunto de prueba

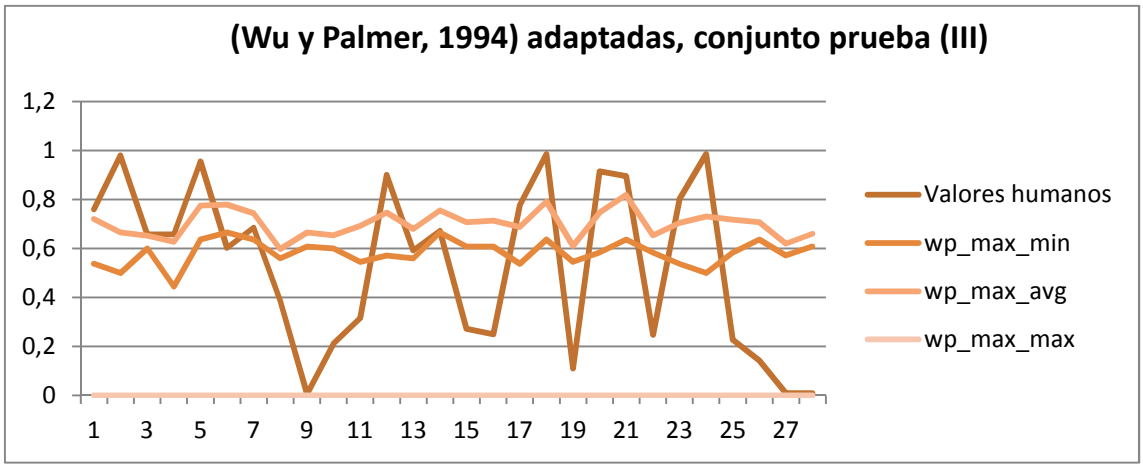

Ilustración 49: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y la medida adaptada de (Wu & Palmer, 1994), que utilizan los valores de profundidad máximos de cada lcs, conjunto de prueba

Calculamos el índice de correlación de Pearson, y comparamos con el valor de la medida original.

Tabla 18: Coeficientes de correlación para las medidas de (Wu & Palmer, 1994) adaptadas, conjunto de prueba

(Wu y Palmer, 1994)	Profundidades	sim_{wp_min}	sim_{wp_avg}	sim_{wp_max}
0,779	Mínima	0,073	0,770	0,801
0,779	Media	0,124	0,709	0,816
0,779	Máxima	0,180	0,617	0,813

Pese a que queda ligeramente por debajo en los valores replicados para el conjunto de entrenamiento (0,787 vs 0,81), supera los resultados en el conjunto de prueba (de 0,779 publicados por (Wu & Palmer, 1994) pasamos a 0,816).

8.2.3. Resultados para la medida de (Leacock & Chodorow, 1994)

Presentamos ahora los valores obtenidos para la medida de (Leacock & Chodorow, 1994) para el conjunto de entrenamiento.

Tabla 19: Valores obtenidos para las medidas de (Leacock & Chodorow, 1994) adaptadas, conjunto de entrenamiento

Concepto1-Concepto2	Valores humanos	Sim_{lc_min}	Sim_{lc_avg}	Sim_{lc_max}
Psychiatric_hospital-Cemetery (1)	0,79	1,29	1,561	1,897
Psychiatric_hospital-Fruit (2)	0,19	1,29	1,616	1,897
Psychiatric_hospital-Monk (3)	0,39	1,386	1,577	1,897
Autograph-Shore (4)	0,06	1,491	1,67	1,897
Autograph-Signature (5)	3,59	1,491	1,845	2,302
Automobile-Magician_(fantasy) (6)	0,11	1,049	1,188	1,29
Automobile-Cushion (7)	0,97	1,203	1,319	1,609
Bird-Woodland (8)	1,24	1,491	1,681	2,079
Boy-Rooster (9)	0,44	1,491	1,685	1,897
Boy-Philosophy (10)	0,96	1,742	1,992	2,079
Cemetery-Mound (11)	1,69	1,29	1,8	2,59
Cemetery-Graveyard (12)	3,88	1,386	1,968	2,995
Cemetery-Woodland (13)	1,18	1,386	1,93	2,995
Rope-Rope (14)	3,41	1,386	1,816	2,995
Rooster-Rooster (15)	3,68	1,203	1,716	2,995
Crane_(bird)-Rooster (16)	1,41	0,98	1,065	1,123
Cushion-Jewellery (17)	0,45	1,491	1,804	2,302
Cushion-Pillow (18)	3,84	1,609	1,747	2,995
Forest-Woodland (19)	3,65	1,386	1,917	3,688
Fruit-Furnace (20)	0,05	1,609	1,886	2,079
Furnace-Tool (21)	1,37	1,29	1,917	2,995
Glass-Jewellery (22)	1,78	1,491	1,589	1,742
Glass-Glass (23)	3,45	1,386	1,757	2,995
Graveyard-Psychiatric_hospital (24)	0,42	1,29	1,561	1,897
Smile-Tool (25)	0,18	1,386	1,616	1,742
Smile-Boy (26)	0,88	1,609	1,782	1,897
Smile-Smile (27)	3,46	1,609	1,967	2,995
Hill-Mound (28)	3,29	1,29	1,485	1,609
Hill-Woodland (29)	1,48	1,386	1,537	1,742
Magician_(fantasy)-Oracle (30)	1,82	1,203	1,54	1,897
Mound-Stove (31)	0,14	1,386	1,546	1,897
Mound-Shore (32)	0,97	1,386	1,591	1,742
Oracle-Philosophy (33)	2,61	1,491	1,982	2,995
Philosophy-Magician_(fantasy) (34)	2,46	1,386	1,688	2,079
Serfdom-Slavery (35)	3,46	1,609	1,988	2,995
Shore-Travel (36)	1,22	1,491	1,67	1,897
Shore-Woodland (37)	0,9	1,491	1,662	1,897

Y a continuación los valores para el conjunto de prueba:

Tabla 20: Valores obtenidos para las medidas de (Leacock & Chodorow, 1994) adaptadas, conjunto de prueba

Concepto1-Concepto2	Valores humanos	Sim_{lc_min}	Sim_{lc_avg}	Sim_{lc_max}
Psychiatric_hospital-Psychiatric_hospital (1)	3,04	1,203	1,657	2,995
Automobile-Automobile (2)	3,92	0,916	1,466	2,995
Bird-Rooster (3)	2,63	1,491	1,589	1,742
Bird-Crane_(bird) (4)	2,63	0,98	1,49	2,59
Boy-Boy (5)	3,82	1,609	1,918	2,995
Sibling-Boy (6)	2,41	1,742	2,318	3,688
Brother_(Catholic)-Monk (7)	2,74	1,491	1,771	2,995
Automobile-Travel (8)	1,55	1,203	1,377	1,609
Rope-Smile (9)	0,02	1,491	1,662	1,897
Coast-Forest (10)	0,85	1,491	1,612	1,742
Coast-Hill (11)	1,26	1,386	1,871	2,995
Coast-Shore (12)	3,6	1,491	1,98	3,688
Crane_(machine)-Tool (13)	2,37	1,123	1,774	2,995
Food-Fruit (14)	2,69	1,609	2,204	3,688
Food-Rooster (15)	1,09	1,491	1,925	2,59
Forest-Graveyard (16)	1	1,491	1,959	2,995
Furnace-Stove (17)	3,11	1,203	1,809	2,59
Jewellery-Jewellery (18)	3,94	1,609	1,853	2,995
Glass-Magician_(fantasy) (19)	0,44	1,29	1,415	1,491
Tool-Tool (20)	3,66	1,203	1,774	2,995
Travel-Travel (21)	3,58	1,609	1,827	2,995
Boy-Magician_(fantasy) (22)	0,99	1,386	1,591	1,742
Magician_(fantasy)-Magician_(fantasy) (23)	3,21	1,203	1,755	2,995
Noon-Noon (24)	3,94	1,203	1,747	2,995
Monk-Oracle (25)	0,91	1,386	1,953	2,995
Monk-Slavery (26)	0,57	1,609	1,85	2,302
Noon-Rope (27)	0,04	1,386	1,449	1,491
Rooster-Travel (28)	0,04	1,491	1,595	1,742

Representamos dentro de una gráfica los valores anteriores, en primer lugar para el conjunto de entrenamiento, y en segundo lugar para el conjunto de prueba.

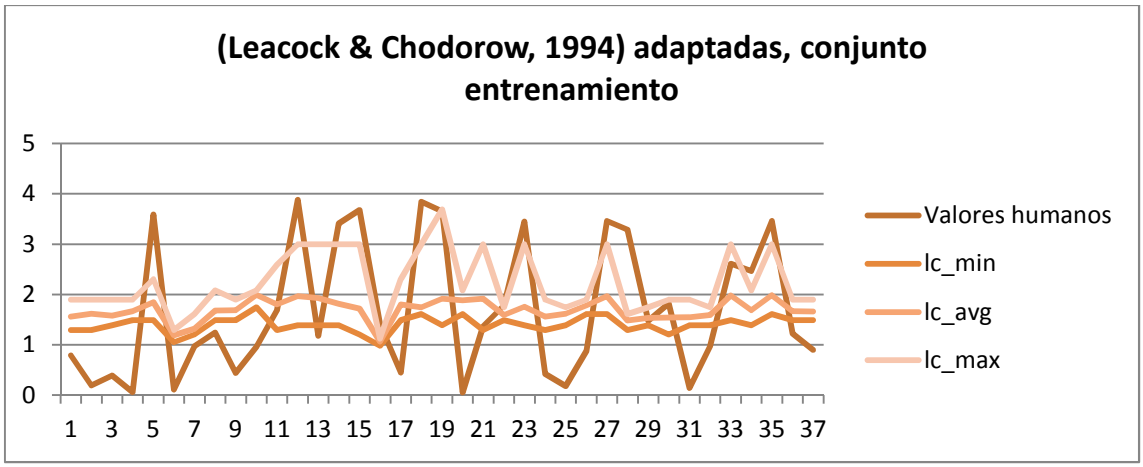

Ilustración 50: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Leacock & Chodorow, 1994), conjunto de entrenamiento

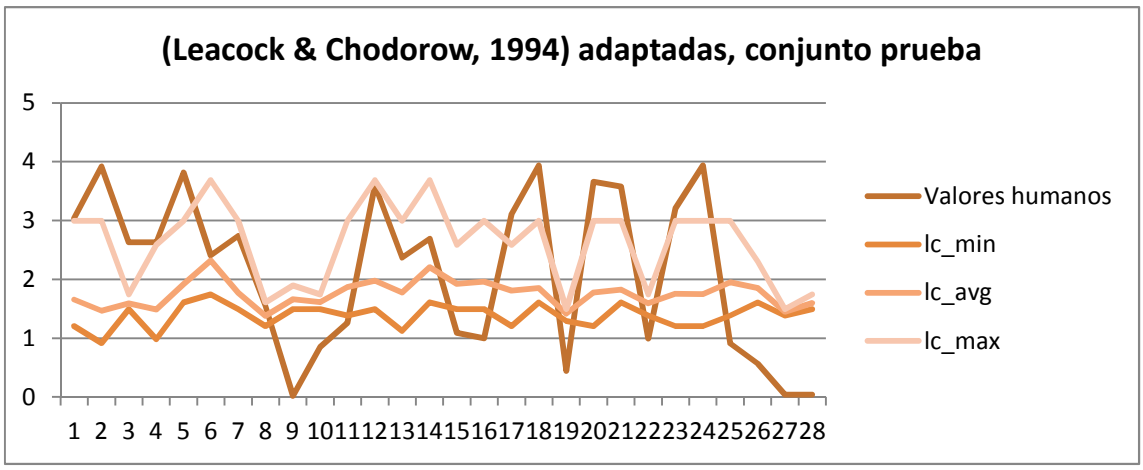

Ilustración 51: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Leacock & Chodorow, 1994), conjunto de prueba

Calculamos ahora el índice de correlación, y comparamos con la correlación de la medida original, para el conjunto de entrenamiento y de prueba (Tabla 21 y Tabla 22, respectivamente).

Tabla 21: Coeficienes de correlación para las medidas de (Leacock & Chodorow, 1994) adaptadas, conjunto de entrenamiento

(Leacock y Chodorow, 1994)	lc_min	lc_avg	lc_max
0,865	0,104	0,402	0,681

Tabla 22: Comparación de coeficientes de correlación para la medida de (Leacock & Chodorow, 1994) para el conjunto de prueba

(Leacock y Chodorow, 1994)	lc_min	lc_avg	lc_max
0,8208	0,208	0,293	0,656

Para esta medida comprobamos que nuestra adaptación se queda bastante por debajo de la medida original, siendo de 0,68 para el conjunto de entrenamiento y de 0,65 para el de prueba los mayores valores obtenidos.

8.2.4. Resultados para la medida de (Blázquez-del-Toro et al., 2008)

El caso de la medida adaptada de (Blázquez-del-Toro et al., 2008) es un poco diferente ya que usa una constante *k* que puede variar entre 0 y 2,5. Nosotros variaremos *k* entre esos valores, y calcularemos las medidas para cada una de las constantes posibles. Posteriormente nos quedaremos con los valores cuya *k* hagan máximo los coeficientes de correlación de los conjuntos de pares de entrenamiento y prueba.

Tabla 23: Valores de los coeficientes de correlación para todos los posibles valores de *k*

	Corr. del conjunto de entrenamiento			Corr. del conjunto de prueba		
K	bl_min	bl_avg	bl_max	bl_min	bl_avg	bl_max
0,25	0,794	0,807	0,795	0,836	0,843	0,848
0,50	0,790	0,805	0,794	0,835	0,842	0,847
0,75	0,785	0,802	0,793	0,834	0,840	0,845
1	0,780	0,798	0,791	0,832	0,838	0,843
1,25	0,774	0,794	0,789	0,830	0,835	0,840
1,50	0,767	0,788	0,784	0,825	0,829	0,834
1,75	0,758	0,780	0,777	0,818	0,821	0,825
2	0,745	0,768	0,765	0,806	0,807	0,809
2,25	0,726	0,748	0,739	0,777	0,772	0,763
2,50	0,692	0,702	0,639	0,629	0,753	0,595

La primera fila corresponde a los valores máximos del coeficiente de correlación. Elegimos la *k* que maximiza esos valores para los dos conjuntos de pares, siendo k=0,25 para los valores mínimos (r=0,794,/r=0,836), para los valores medios (r=0,807/0,843) y para los valores máximos (r=0,795/0,848).

A continuación exponemos los valores de las medidas de Blázquez para los conjuntos de pares de entrenamiento y prueba con k=0,25.

Tabla 24: Valores obtenidos para las medidas de (Blázquez-del-Toro et al., 2008) adaptadas, conjunto de entrenamiento con k=0,25

Concepto1-Concepto2	Valores humanos	bl_min	bl_avg	bl_max
Psychiatric_hospital-Cemetery (1)	0,79	0,022	0,041	0,057
Psychiatric_hospital-Fruit (2)	0,19	0,024	0,470	0,054
Psychiatric_hospital-Monk (3)	0,39	0,020	0,036	0,052
Autograph-Shore (4)	0,06	0,024	0,038	0,052
Autograph-Signature (5)	3,59	0,035	0,057	0,699
Automobile-Magician_(fantasy) (6)	0,11	0,014	0,031	0,042

Automobile-Cushion (7)	0,97	0,018	0,032	0,043
Bird-Woodland (8)	1,24	0,028	0,046	0,061
Boy-Rooster (9)	0,44	0,019	0,404	0,056
Boy-Philosophy (10)	0,96	0,031	0,053	0,070
Cemetery-Mound (11)	1,69	0,047	0,059	0,068
Cemetery-Graveyard (12)	3,88	0,134	0,136	0,137
Cemetery-Woodland (13)	1,18	0,043	0,058	0,072
Rope-Rope (14)	3,41	0,137	0,137	0,138
Rooster-Rooster (15)	3,68	0,137	0,137	0,138
Crane_(bird)-Rooster (16)	1,41	0,020	0,037	0,050
Cushion-Jewellery (17)	0,45	0,027	0,046	0,059
Cushion-Pillow (18)	3,84	0,134	0,136	0,137
Forest-Woodland (19)	3,65	0,126	0,128	0,130
Fruit-Furnace (20)	0,05	0,033	0,047	0,067
Furnace-Tool (21)	1,37	0,049	0,064	0,076
Glass-Jewellery (22)	1,78	0,062	0,074	0,086
Glass-Glass (23)	3,45	0,136	0,137	0,138
Graveyard-Psychiatric_hospital (24)	0,42	0,022	0,041	0,057
Smile-Tool (25)	0,18	0,022	0,039	0,053
Smile-Boy (26)	0,88	0,026	0,047	0,064
Smile-Smile (27)	3,46	0,136	0,136	0,137
Hill-Mound (28)	3,29	0,027	0,037	0,049
Hill-Woodland (29)	1,48	0,028	0,045	0,057
Magician_(fantasy)-Oracle (30)	1,82	0,038	0,053	0,063
Mound-Stove (31)	0,14	0,026	0,038	0,054
Mound-Shore (32)	0,97	0,030	0,040	0,053
Oracle-Philosophy (33)	2,61	0,048	0,064	0,078
Philosophy-Magician_(fantasy) (34)	2,46	0,034	0,062	0,074
Serfdom-Slavery (35)	3,46	0,060	0,077	0,087
Shore-Travel (36)	1,22	0,025	0,042	0,060
Shore-Woodland (37)	0,9	0,034	0,050	0,063

Representamos dentro de una gráfica los valores anteriores para el conjunto de entrenamiento.

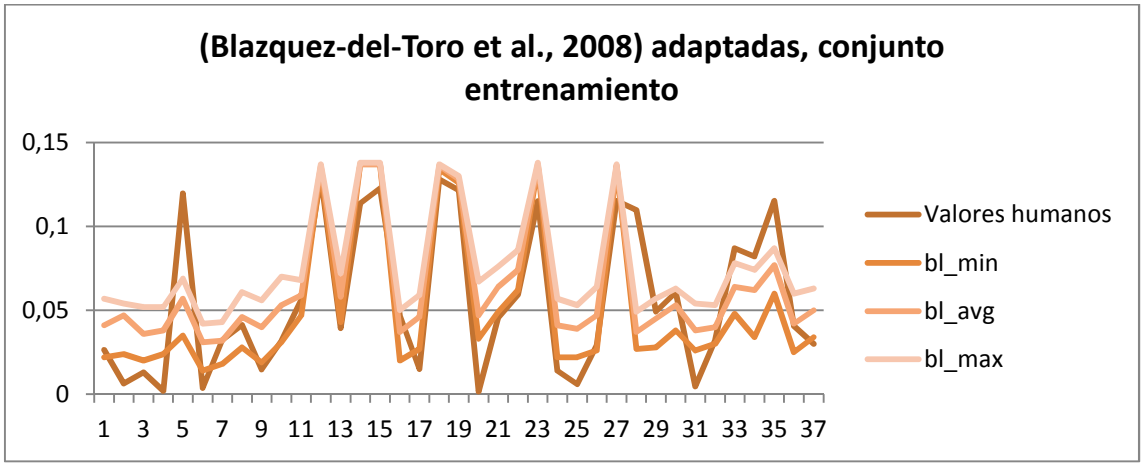

Ilustración 52: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Blázquez-del-Toro et al., 2008), conjunto de entrenamiento con k=0,25

Lamentablemente, la herramienta utilizada para replicar las medidas y obtener los valores del conjunto de entrenamiento, no contiene la medida de (Blázquez-del-Toro et al., 2008), por lo que no incluimos su resultado.

A continuación mostramos los valores de la medida de (Blázquez-del-Toro et al., 2008) adaptada para el conjunto de prueba.

Tabla 25: Valores obtenidos para las medidas de (Blázquez-del-Toro et al., 2008) adaptadas, conjunto prueba con k=0,25

Concepto1-Concepto2	Valores humanos	bl_min	bl_avg	bl_max
Psychiatric_hospital-Psychiatric_hospital (1)	3,04	0,137	0,137	0,138
Automobile-Automobile (2)	3,92	0,138	0,139	0,139
Bird-Rooster (3)	2,63	0,022	0,038	0,055
Bird-Crane_(bird) (4)	2,63	0,060	0,068	0,074
Boy-Boy (5)	3,82	0,135	0,136	0,137
Sibling-Boy (6)	2,41	0,067	0,078	0,092
Brother_(Catholic)-Monk (7)	2,74	0,089	0,094	0,098
Automobile-Travel (8)	1,55	0,019	0,031	0,041
Rope-Smile (9) Chord_(music)	0,02	0,028	0,042	0,053
Coast-Forest (10)	0,85	0,025	0,045	0,058
Coast-Hill (11)	1,26	0,074	0,074	0,082
Coast-Shore (12)	3,6	0,102	0,102	0,108
Crane_(machine)-Tool (13)	2,37	0,065	0,071	0,080
Food-Fruit (14)	2,69	0,058	0,074	0,090
Food-Rooster (15)	1,09	0,041	0,056	0,072
Forest-Graveyard (16)	1	0,044	0,059	0,072
Furnace-Stove (17)	3,11	0,041	0,057	0,068
Jewellery-Jewellery (18)	3,94	0,136	0,136	0,138
Glass-Magician_(fantasy) (19)	0,44	0,018	0,037	0,051
Tool-Tool (20)	3,66	0,136	0,137	0,138
Travel-Travel (21)	3,58	0,136	0,137	0,138
Boy-Magician_(fantasy) (22)	0,99	0,023	0,045	0,062
Magician_(fantasy)-Magician_(fantasy) (23)	3,21	0,111	0,111	0,114
Noon-Noon (24)	3,94	0,137	0,137	0,138
Monk-Oracle (25)	0,91	0,060	0,064	0,073
Monk-Slavery (26)	0,57	0,030	0,047	0,061
Noon-Rope (27)	0,04	0,019	0,039	0,055
Rooster-Travel (28)	0,04	0,024	0,040	0,051

Representamos dentro de una gráfica los valores anteriores para el conjunto de prueba.

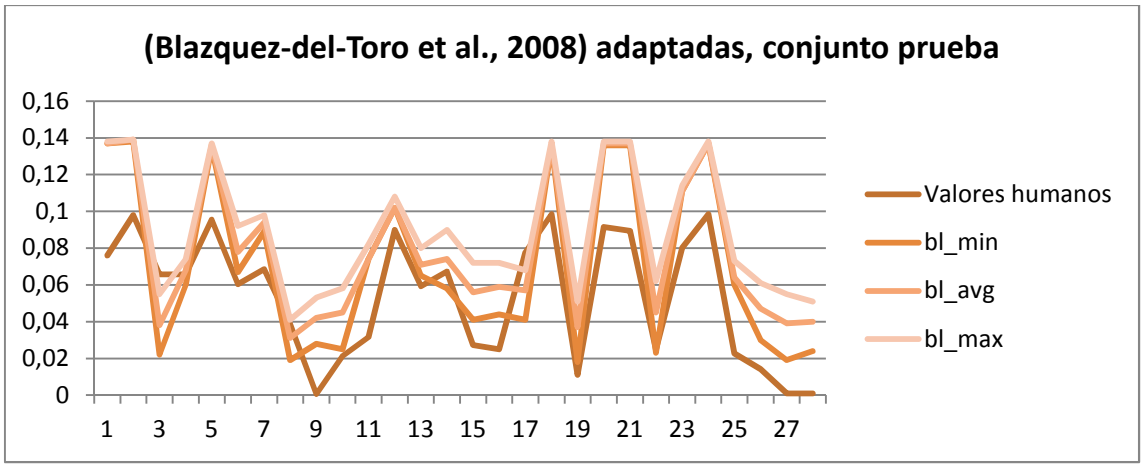

Ilustración 53: Gráfico comparativo de los valores de (Rubenstein & Goodenough, 1965) y las medidas adaptadas de (Blázquez-del-Toro et al., 2008), conjunto de prueba con k=0,25

Calculamos ahora el índice de correlación, y comparamos con la correlación de la medida original.

Tabla 26: Comparación de coeficientes de correlación para las medidas de (Blázquez-del-Toro et al., 2008) adaptadas, conjunto de prueba y k=0,25

(Blázquel-del-Toro et al., 2008)	bl_min	bl_avg	bl_max
0,8269 (k=1,25)	0,836	0,843	0,848

Para esta medida, en el caso del conjunto de prueba supera al índice original (de 0,827 a 0,848).

8.2.5. Resultados para la medida de (Li et al., 2003)

Ahora exponemos los valores para las medidas de (Li et al., 2003). Como ya comentamos, de todos los métodos de (Li et al., 2003), nos quedamos con los cinco que se basan en caminos (enumerándolos del 1 al 5).

La primera medida de (Li et al., 2003) es exactamente igual que la medida de (Rada et al., 1989), por lo que no exponemos los valores de las medidas ya que son exactamente las mismas.

La obtención del resto de medidas de (Li et al., 2003) se realiza con un método diferente del resto de las medidas, ya que habrá que calcularlas en función de constantes. En el caso de (Blázquez-del-Toro et al., 2008), la fórmula de la medida también cuenta con constantes, pero al poder adoptar dichas constantes tan solo diez posibles valores, se calculan todos sus posibles valores para cada par. Para la medida de (Li et al., 2003), se irá procesando para cada par de conceptos y una vez procesados todos los pares, se utilizan

los valores almacenados para calcular el índice de correlación de Pearson para cada valor de α y β comprendidos entre 0 y 1, a intervalos de 0.05, quedándonos con los valores α y β que hacen máximo el índice de correlación, para ambos conjuntos de pares (entrenamiento y prueba).

Exponemos a continuación los valores que maximizan los coeficientes de correlaciones, coincidiendo el valor de las constantes alfa y beta en ambos conjuntos de pares.

Tabla 27: Valores de correlación de Li adaptadas, conjunto de entrenamiento y prueba

	Corr. del conjunto de entrenamiento			Corr. del conjunto de prueba		
Profundidades	**Mínimas**	**Medias**	**Máximas**	**Mínimas**	**Medias**	**Máximas**
Li2min	0,721	0,749	0,760	0,785	0,782	0,790
Li2avg	0,721	0,747	0,760	0,614	0,783	0,758
Li2max	0,720	0,742	0,733	0,785	0,785	0,817
Li3min	0,779	-	-	0,844	-	-
Li3avg	0,779	-	-	0,844	-	-
Li3max	0,779	-	-	0,844	-	-
Li4min	0,779	0,750	0,779	0,844	0,791	0,844
Li4avg	0,779	0,769	0,779	0,844	0,806	0,851
Li4max	0,779	0,774	0,779	0,844	0,823	0,844
Li5min	0,001	0,220	0,733	0,001	0,521	0,719
Li5avg	0,001	0,312	0,671	0,001	0,521	0,608
Li5max	0,438	0,484	0,477	0,001	0,388	0,479

Y en la siguiente tabla los valores de las constantes alfa y beta para obtener dichos resultados.

Tabla 28: Valores de las constantes alfa beta que maximizan los valores de correlación del conjunto de entrenamiento y prueba

	Alfa y beta (alfa en Li3, beta en Li5)		
Profundidades	**Mínimas**	**Medias**	**Máximas**
Li2min	a=0,05 b=1	a=0,3 b=0,65	a=0,25 b=0,65
Li2avg	a=0,05 b=0,85	a=0,45 b=0,55	a=0,25 b=0,85
Li2max	a=1 b=0,05	a=0,9 b=0,65	a=0,5 b=0,75
Li3min	a=0,4	-	-
Li3avg	a=0,4	-	-
Li3max	a=0,4	-	-
Li4min	a=0,4 b=0,15	a=0,3 b=1	a=0,4 b=1
Li4avg	a=0,4 b=0,9	a=0,3 b=1	a=0,35 b=0,2
Li4max	a=0,4 b=1	a=0,35 b=1	a=0,4 b=1
Li5min	b=0,95	b=0,05	b=0,1
Li5avg	b=0,9	b=0,05	b=1
Li5max	b=1	b=0,4	b=0,9

Como podemos observar alcanzamos un valor máximo para el conjunto de entrenamiento de 0,779 y de 0,851 para el de prueba, quedándose ambos valores por debajo de la medida original (0,87 y 0,89 respectivamente).

9. CONCLUSIONES

Este capítulo recoge de modo resumido los aspectos más destacados del proyecto una vez finalizada la construcción del mismo, entre ellos la valoración entre los objetivos que se fijaron inicialmente y lo que realmente se ha conseguido. Además, se discutirá sobre posibles trabajos futuros que podrían llevarse a cabo sobre la aplicación.

9.1. Resultados obtenidos

En el capítulo anterior pudimos comparar nuestros resultados con las medidas tradicionales ya implementadas por otros autores. En la tabla siguiente exponemos un resumen de los valores alcanzados por las medidas originales y los obtenidos de nuestras medidas adaptadas, dentro del conjunto de prueba.

Tabla 29: Valores de las medidas originales y las adaptadas, conjunto de prueba

Medidas de grafos	Valor medida original	Valor medida adaptada, conj. prueba
(Rada et al., 1989)	0,6645	0,785
(Wu & Palmer, 1994)	0,7790	0,816
(Leacock & Chodorow, 1994)	0,8208	0,681
(Blázquez-del-Toro et al., 2008)	0,8269	0,848
(Li et al., 2003)	0,8914	**0,851**

Como podemos ver, aunque algunas no llegan a alcanzar a la medida original, como la medida de (Leacock & Chodorow, 1994), otras igualan o incluso superan la medida original de modo significativo. Por tanto, estas medidas adaptadas constituyen medidas fiables para el cálculo de similitud semántico sobre la fuente Wikipedia; concretamente es la medida adaptada de Li la que mejor resultado da, superando los resultados del resto de medidas originales (exceptuando la del propio (Li et al., 2003) efectuada sobre WordNet). Esta fiabilidad nos permitirá poder desarrollar aplicaciones a partir de la medida de manera adecuada.

Sin embargo, el resultado más interesante del proyecto es poder ver la comparación de nuestro valor máximo obtenido con los conseguidos por las medidas basadas en buscadores Web o Wikipedia. Y es que, como puede verse en la siguiente tabla, esos valores (ya listados en el capítulo 4) son menores que nuestro resultado final.

Tabla 30: Valores de las medidas basadas en buscadores Web junto con nuestra medida

Medidas basadas en buscadores Web	Valor
Bollegala et al., 2007	0,79
WikiRelate!, 2006	0,56
Gabrilovich & Marlovich, 2007	0,75
Wee y Hassen, 2008	0,60
Milne y Witten, 2008	0,64
Zhang et al., 2010	0,56
Nuestra medida	**0,85**

A este hecho hay que añadir la ventaja de tener una fuente de datos como Wikipedia, que como ya comentamos en el estado del arte da la ventaja de ser:

- Uno de los orígenes de datos más grandes y extensibles hoy en día
- Actualizable y actualizado a diario
- Contrastado por una amplia comunidad de usuarios
- Contiene prácticamente casi cualquier concepto existente
- Traducido total o parcial a diferentes lenguajes

A las ventajas obtenidas por el uso de Wikipedia como fuente de datos, incluimos además el procesamiento sencillo que realizamos de su estructura, almacenando sus categorías y relaciones de las mismas. Además, hay que tener en cuenta que, dada la estructura poco jerarquizada (y un poco caótica) de Wikipedia, los resultados obtenidos pueden servir de guía para futuros estudios en estructuras similares. El coste computacional de nuestra medida muy asequible, pues no necesitamos procesar textos como en medidas basadas en corpus.

Se han conseguido los objetivos marcados inicialmente como son, el almacenamiento de la estructura de Wikipedia en base de datos local, y la adaptación de las medidas originales basadas en camino con éxito (alcanzando unas buenas medidas de correlación, siendo incluso superiores a los originales en la mayoría de los casos con respecto al conjunto de prueba, y superiores a las medidas basadas en buscadores Web y Wikipedia.

9.2. Aptitudes adquiridas

Gracias a la realización del proyecto, he adquirido nuevas aptitudes, entre las que puedo enumerar:

- Aprendizaje de un nuevo lenguaje de programación, Ruby.
- Aprendizaje de análisis e implementación de bases de datos mediante MySql
- Aprendizaje de nuevos conocimientos trabajando con programación orientada a objetos.
- Conocimiento global sobre la creación de proyectos y sus fases.
- Conocimientos relacionados con el contenido del proyecto, principalmente medidas de similitud, métricas estadísticas, etc.

9.3. Futuras líneas de desarrollo

A partir del proyecto realizado se pueden crear numerosas líneas de desarrollo que pueden añadir mejoras o ampliaciones sobre lo ya desarrollado. Algunas de ellas podrían ser las siguientes:

- Modificación del código del *Crawler* para que corra en modo *batch* y advertir cambios en Wikipedia: Wikipedia está continuamente actualizándose, por lo que habría que realizar un proceso de control de cambios para anexar la nueva información de modo incremental, y no tener que procesar Wikipedia cada vez que se quiera tener una copia actualizada.
- Mejorar el paso previo de desambiguación de conceptos: automatizar el proceso de desambiguación según ciertos criterios.
- Realizar una interfaz web que permita el uso de nuestra medida por terceras personas de modo on-line.
- Creación de una nueva medida de similitud a partir de valores obtenidos de la estructura de Wikipedia (distancia, profundidad, conectividad, etc.), y que pueda mejorar los resultados obtenidos. De hecho, esta línea se comenzó en el transcurso de este proyecto, consiguiéndose una medida nueva inicial que ofrecía los valores de correlación de 0,81 y 0,86 para los conjuntos de entrenamiento y prueba respectivamente, pero no se ha incorporado su detalle en esta memoria para evitar que su extensión fuese demasiado larga.
- Comprobar si los resultados pueden extrapolarse a otras versiones de diferentes idiomas de Wikipedia.

BIBLIOGRAFÍA

Agirre, & Rigau. (1996). Word sense disambiguation using conceptual density. *Proceedings of the 16th conference on Computational linguistics (COLING '96). Stroudsburg, PA, USA: Association for Computational Linguistics*, (págs. 1, 16-22).

Amescua, Lopez-Cortijo, & García. (1998). Ingeniería del software. *Aspectos de Gestión. Instituto Ibérico de la Industria del Software*.

Blázquez-del-Toro et al. (2008). A semantic similarity measure in the context of semantic queries. *International Journal of Computer Applications in Technology 33, 4*, 285-291.

Bollegala, & et al. (2007). Measuring semantic similarity between words using web search engines. *Proceedings of the 16th international conference on World Wide Web (WWW '07) (Banff, Alberta, Canada, May 8-12)*, (págs. 757-766). New York, NY.

Booch, Rumbaugh, & Jacobson. (2005). *The Unified Modeling Language User Guide, 2/E.* Addison Wesley.

Church, & Hanks. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics 16, 1*, 22-29.

Cilibrasi, & Vitanyi. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering 19, 3*, 370-383.

Gabrilovich, & Marlovich. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence.* Hyderabad, India.

Hassan, & Wee. (2008). Exploiting Wikipedia for Directional Inferential Text Similarity. *Fifth International Conference on Information Technology: New Generations (ITNG 2008).*

Jiang, & Conrath. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Conf. on Research in Computational Linguistics.*

Kozima, & Furigori. (1993). Similarity between words computed by spreading activation on an english dictionary. *the sixth conference on European chapter of the Association for Computational Linguistics*, (págs. 232-239).

Larman, C. (2005). *Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development.* Prentice Hall Professional.

Leacock, & Chodorow. (1994). *Filling in a sparse training space for word sense disambiguation.* ms.

Lee, & et al. (1993). Information Retrieval Based on Conceptual Distance in IS-A Hierarchies. *Journal of Documentation, Vol. 49, No. 2.* , 188-207.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation (SIGDOC '86)*, (págs. 24-26). Toronto, Ontario, Canada.

Li et al. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering 15, 4*, 871-882.

Lin, & et al. (1998). An Information-Theoretic Definition of Similarity. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*, (págs. 296-304). San Francisco, CA, USA.

Miller, & Charles. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes 6, 1*, 1-28.

Milne, & Witten. (2008). An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08).* Chicago, USA.

Rada et al. (1989). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics 19, 1*, 17-30.

Resnik. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th international joint conference on*

Artificial intelligence (págs. 448-453). Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc, San Francisco, CA, USA.

Richardson, & Smeaton. (1995). Using WordNet in a Knowledge-Based Approach to Information Retrieval. *CA-0395*.

Rubenstein, & Goodenough. (1965). Contextual correlates of synonymy. *Communications of the ACM 8, 10*, 627-633.

Strube, & Ponzetto. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI'06)* (págs. 1419-1424). Boston, Massachusets: AAAI Press, Menlo Park.

Sussna. (1993). Word sense disambiguation for free-text indexing using a massive semantic network. *Proceedings of the second international conference on Information and knowledge management (CIKM-93)*, (págs. 67-74). Washington, D.C., USA.

Trillo, & et al. (2007). Discovering the Semantics of Keywords: An Ontology-based Approach. *The 2006 International Conference on Semantic Web and Web Services (SWWS'06).* Las Vegas.

Wilks, & et al. (1990). Providing machine tractable dictionary tools. *Machine Translation 5, 2*, 99-154.

Wu, & Palmer. (1994). Verbs semantics and lexical selection. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics (ACL '94)* (págs. 133-138). Las Cruces, New Mexico: Association for Computational Linguistics, Stroudsburg, PA, USA.

Zhang et al. (2011). Mining and Explaining Relationships in Wikipedia. *IEICE TRANSACTIONS on Information and Systems Vol.E95-D No.7*, 1918-1931.

Tabla 31: 28 parejas de términos y sus conceptos desambiguados

Par de palabras originales		Par de conceptos en Wikipedia	
Asylum	Madhouse	Psychiatric_hospital	Psychiatric_hospital
Automobile	Car	Autmobile	Automobile
Bird	Cock	Bird	Rooster
Bird	Crane	Bird	Crane_(bird)
Boy	Lad	Boy	Boy
Brother	Lad	Sibling	Boy
Brother	Monk	Brother_(Catholic)	Monk
Car	Journey	Automobile	Travel
Chord	Smile	Chord_(music)	Smile
Coast	Forest	Coast	Forest
Coast	Hill	Coast	Hill
Coast	Shore	Coast	Shore
Crane	Implement	Crane_(machine)	Tool
Food	Fruit	Food	Fruit
Food	Rooster	Food	Rooster
Forest	Graveyard	Forest	Graveyard
Furnace	Stove	Furnace	Stove
Gem	Jewel	Gemstone	Gemstone
Glass	Magician	Glass	Magician_(fantasy)
Implement	Tool	Tool	Tool
Journey	Voyage	Travel	Travel
Lad	Wizard	Boy	Magician_(fantasy)
Magician	Wizard	Magician_(fantasy)	Magician_(fantasy)
Midday	Noon	Noon	Noon
Monk	Oracle	Monk	Oracle
Monk	Slave	Monk	Slavery
Noon	String	Noon	Rope
Rooster	Voyage	Rooster	Travel

Tabla 32: 37 parejas de palabras (de Rubenstein y Goodenough) desambiguados

Par de palabras originales		Par de conceptos en Wikipedia	
Asylm	Cemetery	Psychiatric_hospital	Cemetery
Asylum	Fruit	Psychiatric_hospital	Fruit
Asylum	Monk	Psychiatric_hospital	Monk
Autograph	Shore	Autograph	Shore
Autograph	Signature	Autograph	Signature
Automobile	Wizard	Automobile	Magician_(fantasy)
Automobile	Cushion	Automobile	Cushion
Bird	Woodland	Bird	Woodland
Boy	Rooster	Boy	Rooster
Boy	Philosophy	Boy	Philosophy
Cemetery	Mound	Cemetery	Mound

Cemetery	Graveyard	Cemetery	Graveyard
Cemetery	Woodland	Cemetery	Woodland
Cord	String	Rope	Rope
Cock	Rooster	Rooster	Rooster
Crane	Rooster	Crane_(machine)	Rooster
Cushion	Jewell	Cushion	Gemstone
Cushion	Pillow	Cushion	Pillos
Forest	Woodland	Forest	Woodland
Fruit	Furnace	Fruit	Furnace
Furnace	Implement	Furnace	Tool
Glass	Jewel	Glass	Gemstone
Glass	Tumbler	Glass	Glass
Graveyard	Madhouse	Graveyard	Psychiatric_hospital
Grin	Implement	Smile	Tool
Grin	Lad	Smile	Boy
Grin	Smile	Smile	Smile
Hill	Mound	Hill	Mound
Hill	Woodland	Hill	Woodland
Magician	Oracle	Magician_(fantasy)	Oracle
Mound	Stove	Mound	Stove
Mound	Shore	Mound	Shore
Oracle	Sage	Oracle	Philosophy
Sage	Wizard	Philosophy	Magician_(fantasy)
Serf	Slave	Serfdom	Slavery
Shore	Voyage	Shore	Travel
Shore	Woodland	Shore	Woodland